University of Huddersfield Repository

Wang, Jing

Spatio-Temporal Volume-based Video Event Detection

Original Citation

Wang, Jing (2012) Spatio-Temporal Volume-based Video Event Detection. Doctoral thesis, University of Huddersfield.

This version is available at http://eprints.hud.ac.uk/id/eprint/17552/

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

http://eprints.hud.ac.uk/

SPATIO-TEMPORAL VOLUME-BASED VIDEO EVENT DETECTION

JING WANG

A thesis submitted to the University of Huddersfield

in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

School of Computing and Engineering

University of Huddersfield

February 2012

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright") and he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

II. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

III. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

Acknowledgements

First and foremost, I would like to thank the School of Computing and Engineering at the University of Huddersfield for providing this great opportunity to undertake this research with their continuous support to the project and myself.

I would like to thank my first supervisor, my director of studies, Dr. Zhijie Xu. With his great help during my research, I have changed from a young student into a researcher. He gives me many suggestions, fresh ideas, and an ideal experiment environment in the Computer Graphics, Imaging and Vision Research Group.

I would also like to thank all my friends, family and so many excellent colleagues for all their support during my postgraduate study.

Abstract

Online and offline video clips provide rich information on dynamic events that occurred over a period of time, for example, human actions, crowd behaviours, and other subject pattern changes. Although substantial progresses have been made in the last 3 decades on 2D image feature processing and their applications in areas such as face matching and objects recognition, video event detection still remains one of the most challenging fields in computer vision study due to the wide range of continuous and non-linear signals engaged by an imaging system, and the inherent semantic difficulties in machine-based understanding of the detected feature patterns.

For bridging the gap between the pixel-level image features and the semantic "meanings" of a videoed single human event, this research has investigated the problem domain through employing the 3D Spatio-Temporal Volume (STV) structure and its global feature paradigm for event pattern recognition. The process pipeline follows an improved Pair-wise Region Comparison (I-PWRC) and a region intersection (RI) based 3D template matching approach for detecting and identifying human actions under uncontrolled real-world videoing conditions. To maintain the run-time performance of this innovative system design, this programme has also developed an efficient pre-filtering mechanism to reduce the amount of voxels (volumetric pixels) that need to be processed in each operational cycle.

For further improving the system's adaptability and robustness, several optimisation techniques, such as scale-invariant template matching and event location prediction mechanisms, have also been developed and implemented. The proposed design has been tested on various renowned online computer vision research databases and been benchmarked against other classic implementation strategies and systems. Satisfactory evaluation results have been obtained through statistical analyses on standard test criteria such as "Recall" rate and the processing efficiency.

List of Publications

[1] Jing Wang, Zhijie Xu. STV Feature Processing for Video Event Detection, Signal Processing, 2011, Submitted.

[2] Jing Wang, Zhijie Xu. Video Event Detection Based on Over-segmented STV Regions. In: 2011 IEEE International Conference on Computer Vision Workshops. Barcelona, Spain. 2011, pp. 1464-1471. ISBN 9781467300636

[3] Jing Wang, Zhijie Xu, Video Analysis Based on Volumetric Event Detection. International Journal of Automation and Computing, 7 (3), 2010, pp. 365-371. ISSN 1476-8186.

[4] Jing Wang, Zhijie Xu, Michael O'Grady. Head Curve Matching and Graffiti Detection. In: British Machine Vision Conference 2010. Aberystwyth, UK. 2010

[5] Jing Wang, Michael O'Grady, Qian Xu, Zhijie Xu. Video Feature Representation and 3D Curve-based Event Matching. In: 16th International Conference on Automation and Computing (ICAC'10), University of Birmingham. UK, 2010.

[6] Jing Wang, Zhijie Xu, Jonathan Pickering. Volume-based Video analysis using 3D Segmentation Techniques. In: 15th International Conference on Automation & Computing. Luton, UK. 2009.

[7] Jing Wang, Zhijie Xu, Qian Xu. Video Volume Segmentation for Event Detection. In: Computer Graphics, Imaging & Visualization, new advances and trends. IEEE Computer Society, Tianjin, China, pp. 311-316, 2009.ISBN 9780769537894

[8] Jing Wang, Zhijie Xu, Jonathan Pickering, Rakesh Mishra. An innovative volume based video feature extraction technique. In: Proceedings of Computing and Engineering Annual Researchers' Conference 2008: CEARC'08. University of Huddersfield, Huddersfield, pp. 110-116. 2008. ISBN 9781862180673

List of Symbols & Abbreviations

AC	Active Contour
AI	Artificial Intelligent
AUC	Area under Curve
BoW	Bag-of-Words
CBIR	Content-based Image Retrieval
CCD	Charge-Coupled Device
CCTV	Closed Circuit Television
CF	Coefficient Factor
CMOS	Complementary Metal–Oxide–Semiconductor
CT	X-ray Computed Tomography
CUDA	Compute Unified Device Architecture
CV	Computer Vision
DIP	Digital Image Processing
DoG	Difference-of-Gaussian
DoK	Dictionary of Keys
FBF	Frame-by-Frame
FIFO	First-In-First-Out
FPS	Frames per Second
GF	Gate Function
GMI	Geometric Moment Invariants
GPU	Graphic Process Unit
HCI	Human Computer Interaction
HMM	Hidden Markov Model
IA	Interest Area
I-PWRC	Improved Pair-wise Region Comparison
LPP	Locality Preserving Projections
LUT	Look Up Table
MRI	Magnetic Resonance Imaging
MS	Mean Shift
MST	Minimum Spanning Tree
*n*D	*n*-dimensional
OCR	Optical Character Recognition
PCB	Printed Circuit Board
PDF	Probability Density Function
PRC	Precision Recall Curve
PWRC	Pair-wise Region Comparison
RI	Region Intersection
ROC	Receiver Operator Characteristic
ROI	Regions of Interest
SIFT	Scale Invariant Feature Transform
STV	Spatio Temporal Volume
ToF	Time-of-Flight

List of Figures

Figure 1-1 CV Research hotspots timeline 3

Figure 1-2 CV Research framework 6

Figure 2-1 "Ω"contours and matching result 12

Figure 2-2 Definition of Spatio-temporal Volume 14

Figure 2-3 STV model construction operation 16

Figure 2-4 Different segmentation outputs based on same low-level features but .different applications 17

Figure 2-5 Artificial images contains gradients, solid and noise area 18

Figure 2-6 Template matching system pipeline 22

Figure 2-7 Machine learning system pipeline 22

Figure 2-8 (a) "Waving" template, (b), (c), (d) selected snapshots of detected events 23

Figure 2-9 Application-specific "fall down" event 25

Figure 2-10 STV slices used for human gait analysis 29

Figure 2-11 System hierarchical structure 32

Figure 3-1 Segmentation pipeline used in this research 35

Figure 3-2 Segmentation output of Figure 2-5 by using newly developed approach introduced in this research 36

Figure 3-3 Snapshots of PWRC segmentation outputs 40

Figure 3-4 Factor *k* is not flexible to deal with complex real-world environment 41

Figure 3-5 Graph formed by using region histogram 46

Figure 3-6 I-PWRC Hierarchical graph representation 48

Figure 3-7 Snapshots of video clips for I-PWRC feasibility test 50

Figure 3-8 Sample segmentation outputs for original video inputs 51

Figure 3-9 Frame comparison between baseline PWRC and I-PWRC 53

Figure 4-1 A "Falling down" event represented in STV model 56

Figure 4-2 Building a "Waving" event template requires FBF-based RoI operations 58

Figure 4-3 The illustration of the active contour algorithm 60
Figure 4-4 Active Contour-based “Waving Template” Formation 61
Figure 4-5 RI template matching algorithm and four possible scenarios 64
Figure 4-6 Region filtering pipeline 68
Figure 4-7 Binary form STV Shape of a waving template 68
Figure 4-8 Local histograms used for evaluating the intersected regions 70
Figure 4-9 A Local histogram registering a “perfect” matching 71
Figure 5-1 System pipeline 74
Figure 5-2 Case for STV normalisation 75
Figure 5-3 SIFT process flowchart 79
Figure 5-4 SIFT candidate comparison 80
Figure 5-5 SIFT Features 80
Figure 5-6 3D Interest area construction 82
Figure 5-7 Efficiency improvement from interest area identification 83
Figure 5-8 System hardware platform 83
Figure 5-9 Volume buffer procedures 86
Figure 5-10 The flowchart of the MS pre-segmentation algorithm 87
Figure 5-11 Constructing histogram-based region graph for I-PWRC 89
Figure 5-12 Region Filtering algorithm 91
Figure 5-13 Processes for filtering out the “non-contributing” sub-regions 91
Figure 5-14 STV-based RI matching algorithm 92
Figure 6-1 Snapshots from Weizmann datasets 95
Figure 6-2 Snapshots from KTH datasets 96
Figure 6-3 Snapshots from the self-made Campus datasets 96
Figure 6-4 Artificial event model and STV hierachical strcture 99
Figure 6-5 Improvements on time consumptions from the “Filtering” and “RI Matching” phases 99
Figure 6-6 The KTH confusion matrix 101

Figure 6-7 Template matching results (ROC and RP curves) on the campus dataset..... 103

Figure 6-8 MS over-segmentation result (h_c=5, h_l=5) on 3 events..... 105

Figure 6-9 MS segmentation results (h_c=15, h_l=15) on 3 events..... 106

Figure 6-10 MS and I-PWRC based RI matching confusion matrices 107

Figure 6-11 KTH datasets average detection accuracy based on different colour spaces. 109

Figure 6-12 Accuracy impact of average template numbers..... 110

Figure 6-13 KTH confusion matrices after employing the using multi-scaled templates. 111

Figure 6-14 Normalised multi-scaled templates applied on the Campus dataset..... 112

Figure 6-15 Selected frames from a “people falling down” video clip..... 113

Figure 6-16 Over-segmented sub-regions from the “falling down” event 114

Figure 6-17 Performance comparisons between classic approaches..... 114

List of Tables

Table 2-1 Application domains and performance demands 33

Table 5-1 Relations between the thresholds and the number of templates 76

Table 5-2 Volume buffer Pseudo code 86

Table 5-3 Pseudo code for the I-PWRC method..... 90

Table 6-1 Selected datasets used for evaluations..... 95

Table 6-2 Confusion matrix 97

Table 6-3 Parameters used for KTH Dataset 100

Table 6-4 Matching accuracy performance compared with other approaches..... 101

Table of Contents

Copyright Statement I

Acknowledgements II

Abstract III

List of Publications IV

List of Symbols & Abbreviations V

List of Figures VI

List of Tables IX

Table of Contents X

Chapter 1. Research Background 1

1.1. Brief Note on Computer Vision 2

1.2. Research Framework 5

1.3. Contribution to Knowledge 7

1.4. Thesis Organisation 8

Chapter 2. Related Works 9

2.1. Image and Video Feature 9

2.1.1. Feature Points and Feature Space 9

2.1.2. Feature Definition Strategies 10

2.2. Spatio-temporal Volume Model 13

2.3. Segmentation Methodologies 16

2.3.1. Image Segmentation for Feature Extraction 17

2.3.2. 2D Segmentation Strategies 18

2.4. Pattern Recognition Approaches 21

2.5. Video Event Definition and Human Action Detection 23
2.5.1. Current Video Event Detection Research and Practices 25
2.5.2. Reviewing of STV-based Event Detection 27
2.6. Application Domains 32
2.7. Prominent Challenges for Video Event Detection 34
Chapter 3. Segmentation-based Shape Feature Extraction 35
3.1. Baseline Methods for STV Feature Extraction 36
3.1.1. Baseline 2D PWRC 37
3.1.2. I-PWRC for STV Segmentation 40
3.2. I-PWRC Implementation 42
3.2.1. Pre-clustering Using Mean Shift Algorithm 42
3.2.2. Histogram-based region description 45
3.2.3. Hierarchical Pair-wise Region Comparison 48
3.3. Feasibility Studies 49
3.4. Summary 54
Chapter 4. Volumetric Shape Extraction for Event Template Matching 55
4.1. Event Template Definition 56
4.1.1. Template Matching Strategy Design 56
4.1.2. Forming Event Template 57
4.1.3. AC Concepts 59
4.1.4. AC Implementation Principles 60
4.2. Event Shape Matching 62
4.2.1. Practical Issues 62
4.2.2. Region Intersection Strategy 63
4.3. Renovating the RI method 66

4.3.1. Improved Region Filtering....66
4.3.2. Histogram-Verified Coefficient Factors....69
4.4. Summary and Discussions....72
Chapter 5. Implementation Strategy and System Prototyping....73
5.1. STV Normalisation....74
5.1.1. Hierarchical and Multi-scaled Templates....75
5.1.2. Normalising the Multi-scaled Templates....77
5.2. RI Interest Area Identification....77
5.2.1. Locating the Interest Feature Points....78
5.2.2. SIFT-based Interest Area Formulation....81
5.3. System Modularisation and Data Pre-processing....83
5.3.1. Data Filtering Consideration....84
5.3.2. STV Buffering Technique....85
5.4. STV Feature Extraction....87
5.5. Event Shape Matching....90
5.6. Summary and Discussion....93
Chapter 6. Experiment and Evaluation....94
6.1. Test Data Acquisition....94
6.2. System Performance Evaluations....97
6.2.1. Evaluation Benchmark....97
6.2.2. Efficiency Measurements....98
6.2.3. Matching Accuracy Evaluations....100
6.2.4. Model Compatibility Evaluations....103
6.2.5. Matching Performance within Different Colour Spaces....108
6.2.6. Template representation and Matching Performance....109

6.3. Scale-Invariant Event Detection 110
6.4. Test on Uncontrolled Video Inputs 112
Chapter 7. Conclusions and Future Work 115
7.1. Programme Summaries 115
7.1.1. I-PWRC Validity and Practicality 115
7.1.2. 3D RI Matching Adaptability and Robustness 117
7.1.3. Function Modularisation and System Integration 119
7.2. Future Work 120
References 124

Chapter 1. Research Background

After 40 years of intensive research on image and video processing in academic societies, a wide spectrum of Computer Vision (CV) techniques and applications have been deeply permeated into people's daily lives, for example, Closed Circuit Television (CCTV) surveillance systems, traffic control deployments, and medical and scientific imaging. By using ever more affordable videoing devices such as the digital camera, DV recorder and even the mobile phone, video data is becoming a popular media form stored and circulated over the Internet, as evidenced by the claim that more than 20 minutes of video clips are uploaded to various websites every second [Jarrett 2010]. In addition, CCTV surveillance systems have been increasingly installed in public areas for crime prevention, crowd control, and emergency management. A survey scheme carried out in 2011 reported that an estimation of 1.85 million CCTV cameras existed in the UK [Fry 2011]. It is anticipated that accompanied by the advancement and the application trend of video formats and videoing devices, future video data processing and analysis technologies will have increased demands on effectiveness, automation, intelligence, and timeliness.

However, compared with the fundamental video processing areas such as digitisation, encoding and compressing, higher level applications like vision-based video event detection are still facing many challenges, including video feature segmentation, classification, noise removal, tracking, and pattern recognition, which are often hampered by problems such as video quality, lighting and occlusion.

In this research, an improved video event detection approach has been investigated to tackle part of the problems mentioned above through using various video assemblies,

feature extraction, and template matching algorithms. The research started from defining single human actions in the so-called Spatio-Temporal Volume (STV) space that was first proposed by [Adelson and Bergen 1985] to represent the global spatial and temporal information in a video. The STV and its corresponding data structures enable an event matching task to be transformed into a 3D model comparison and analysis operation. It is evident in this research that through appropriate transformations, static and dynamic information can be encapsulated into corresponding 3D "shapes", whose envelopes and internal compositions can be extracted and studied in the global feature space. It is also observed in the project that many 2D pattern recognition techniques can be readily extended into the 3D volumetric domain for template-based matching processes. The main objective of this thesis is to record and discuss the key research findings as well as the corresponding development and evaluation.

1.1. Brief Note on Computer Vision

The very first question referred to as computer vision (CV), "how to build an artificial vision system for guiding robots to perform the same functions as human vision system does?" [Boden 2006] has been an extremely challenging one since the early stage of Artificial Intelligent (AI) research in the 1970s. As a key component to the modern AI-driven robots, visual cues are defined and acquired as special 2D signals from optical sensors, which share the same technological foundations as signals generated by other acoustic, pressure and temperature sensors. However, comparing with those conventional sensor signals, the real challenge is that vision signals

contains more than 80% information-entropy [Davison 2005], which can be explored to a great depth-evidence to the maxim "a picture is worth a thousand words".

For tackling the CV problems, many research questions have been introduced during its development. The hot-spots in chronological order on computer vision can be illustrated in Figure 1-1.

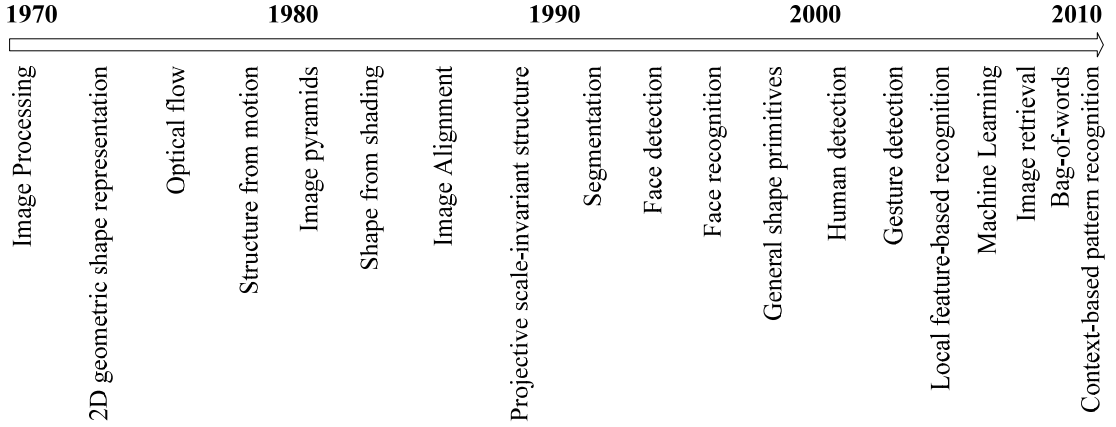


Figure 1-1 CV Research hotspots timeline

As shown in the figure, some of the early pioneering work had been focusing on image signal processing and optimisation strategies for enabling high-level image understanding. Being rooted to the traditional signal processing domains, such as filtering and frequency feature analysis, those researchers have facilitated a development in image quality enhancement, feature detection and camera calibration. Almost as a by-product, some research and pilot projects have started paying attention to video processing through adopting 2-dimensional (2D) image processing techniques onto the consecutive video frames [Wang and Cohen 2007].

Around the 1980s, many mathematical models for tackling image feature representation and content analysis problems were investigated such as [Katsuragawa *et al* 1988], [Yuille and Grzymacz 1988] and [Feddema *et al*. 1989]. These mathematical theories and their applications have provided a blueprint for the general

pattern recognition system structures that are essential for driving practical vision applications. Foundational methodologies and techniques, such as 3D reconstruction [Marin 1987] and shape from motion [Terzopoulos *et al.*1988] are among the prominent research landmarks from this work.

Based on the developments in mathematical modelling and the fast growing processing speed of computer hardware in 1990s, research on computer vision techniques for solving real-world application problems became popular. One of the most spectacular achievements in modern computer vision research came from face recognition, which was actually the first time that a computer vision application can be declared as a real AI system. The developments on face recognition across the entire 90s had also improved the corresponding image segmentation, feature point extraction, and statistical pattern analysis methodologies.

Entering the new millennium, motivated by the great success in face recognition, a number of leading CV research group have been studying problems such as gesture and action recognition [Bobick and Davis 2001], crowd behaviours [Zhao and Nevatia 2002] and motion prediction [Bommer 2005]. Valid research outcomes on human detection, tracking, and action recognition methods have been applied in interpreting the semantic meanings involved in images and videos, for example, global and local feature based human action representation, skeleton model reconstruction, event template matching, and machine learning algorithms-based feature categorisation.

Recently, the CV research has been experiencing a weight shift to the semantic-informed application domains, for example, social-network-based online image/video retrieval [Stone *et al.* 2008], natural language assisted interactive video illustration [Fathi *et al.* 2011], and other diversified areas such as sparse feature representation

[Wright *et al.* 2009]. It is envisaged that with ever more powerful computers and digital image and video equipment, the demands for portable and real-time computer vision technologies and services will increase dramatically, which present both challenges and opportunities for the researchers in the field.

1.2. Research Framework

From the viewpoint of process modularisation, computer vision application systems often follow a process pipeline composed of three modules, data acquisition, feature extraction, and pattern recognition as illustrated in Figure 1-2. Each module can also be further divided into detailed operations depending on the deployed processing strategy and algorithms. For example, data acquisition often encompasses the stages from receiving sensory signals to data pre-processing such as filtering and decompressing.

At the front-end of the framework, image capturing techniques play an important role in building a firm foundation for the entire vision system. For example, by using state-of-the-art depth sensors [Foix *et al*. 2011] or multi-view point-based image capturing methods [Agarwala *et al.* 2006], 3D scenes can be reconstructed intuitively that save precious computational time for other maths-intensive and pattern-based operations.

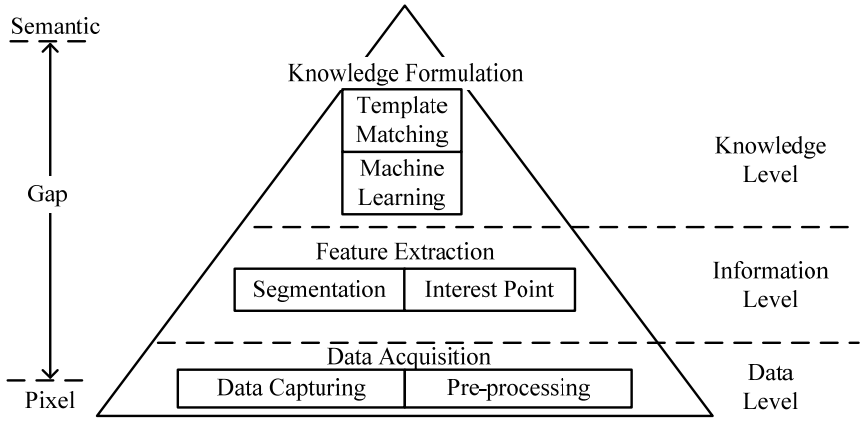


Figure 1-2 CV Research framework

Following the image capturing steps, vision signals are treated in a pre-processing module. Specific filtering techniques, such as compression [Sayood 1996], and de-noise [Jong-Sen 1981] are involved in this step, which provide reliable datasets for the following feature extraction and pattern recognition steps.

Based on the classic concept in Information Theory [Davison 2005], data needs to be processed and refined to generate information before knowledge can be abstracted and modelled from the information. The refinement and categorising works from data to information is realised in this pipeline through the feature extraction module. Generally speaking, an image feature is a mathematical representation of certain image characters bodies. For example, the skin colour model of human bodies can be represented by pixel groups acquired by image segmentation techniques introduced by [Yang and Ahuja 1998].

The tail-end of the CV research framework is the pattern recognition module, which translates the mathematical models into semantic descriptions of the subject that people can understand- from information to knowledge. Recent developments in this field have been focused on geometric-based methods such as model fitting [Hothorn

et al. 2010], contouring and aspect graphs [Giorgi *et al.* 2010]. Other probabilistic and inferential methods have also proven beneficial for this purpose.

A comprehensive coverage of the entirety of the computer vision research framework is beyond the scope of this thesis. The video event detection methodology and the system prototype developed in this research had been focusing on the feature extraction and recognition modules to highlight the validity and robustness improvements of the proposed approach.

1.3. Contribution to Knowledge

The main contributions of this research are summarised as follows:

- An innovative 3D volumetric segmentation method has been devised and implemented based on Graph Theory. When applied in segmenting 3D volumetric event models, the new approach generated superior effectiveness and efficiency over other conventional 2D-based image segmentation methods (Section 3.1 and 3.2).

- Based on the improved STV feature sets, a volumetric shape matching algorithm has been developed and is capable of handling action events recorded in noisy real-world conditions. (Section 4.2 and 4.3).

- The successful integration of feature point and feature-region-based template matching approaches to harness and augment the advantages from both the local and global feature domains (Section 5.1and 5.2).

1.4. Thesis Organisation

This dissertation is arranged in the following order: Chapter 1 and 2 presents a comprehensive review of the project background and the state-of-the-art of the research domain. Chapter 3 focuses on the STV feature extraction techniques developed in the project. The developed STV pattern recognition method and the corresponding event template definition strategies have been introduced in Chapter 4. Chapter 5 reports the system integration and quality reassurance developed and deployed in the process pipeline. In Chapter 6, a series of quantitative experiments have been carried out and result have been analysed for evaluation. Chapter 7 concludes the research with envisaged future works.

Chapter 2. Related Works

Video event detection research encompasses a wide spectrum of studies on Digital Image Processing (DIP), video compressing, pattern analysis and even biological vision. Since its birth in the 1970s, the research outcomes have had impact on many fronts with extensive applications found in industry, such as traffic monitoring systems, CCTV-based security and surveillance networks, and robotic control. This chapter focuses on the prominent work to date on video event analysis and STV-based template matching, which starts with a detailed introduction on related and essential background knowledge. The Section 1 introduced fundamental techniques used in this research. Section 2 and 3 covered the feature definition and extraction strategies. Section 4 reviewed the methodologies used for the pattern recognition. Section 5 highlighted the video event detection framework proposed in this project. The application and challenges involved in the related research area were introduced in Section 6 and 7, consecutively.

2.1. Image and Video Feature

2.1.1. Feature Points and Feature Space

In image processing and pattern recognition, the concepts of “feature” and “feature space” are commonly used. Generally speaking, features are representations of image “signals” at information level as one or multiple mathematical models for further analysis. Features can be extracted from pixels, the fundamental elements of image, by using specific representation methods based on different applications.

For example, in a STV model, a voxel **v**, can be represented as a 6-dimension (6D) vector containing the location and colour information as denoted in Equation 2-1

$$\mathbf{v} = [x, y, z, r, g, b], \quad \textbf{2-1}$$

where *x*, *y*, *z* denotes the coordinates of the voxel, and *r*, *g*, *b* denotes the red, green and blue colour values.

A feature space is an *n*-dimensional coordinate system containing pre-defined feature "points". Based on the complexity and entropy of extracted information, different image features and corresponding spaces can be divided into a two-level hierarchical structure: the low-level feature space contains feature points which can be directly abstracted from pixel/voxel and their neighbours based on their values and coordinates. The high-level features are built upon the low-level contents and defined based on semantic information. Actually, the accurate extraction and representation of high-level features are one of the research hot-spots in pattern recognition domain. Related research works such as Content-based Image Retrieval (CBIR) [Datta *et al.* 2008], and texture-based image segmentation [Jain and Farrokhnia 1991] have been applied in many successful applications.

2.1.2. Feature Definition Strategies

Since the very beginning of the pattern recognition research in 1930s [Duda and Hart 1973], abstracting appropriate features for pattern analysis has been a popular research area. Generally speaking, most matured feature extraction techniques can be classified into three categories: geometric features, statistical features and dynamic features.

- Geometric Features

Geometric features are relatively easy to be defined and extracted from an image or a video frame. It often serves as the fundamental material for more complicated feature categories. One representative application of using geometric feature is the Optical Character Recognition (OCR) System [Mori *et al.* 1999]. The success of this application is one of the most significant milestones of the pattern recognition research of the 1990s and has seen many real-world adoptions in, e.g. hand writing recognition, and new Human Computer Interaction (HCI) design.

Among the popular geometric features, the so-called Geometric Moment Invariants (GMI) feature, which was first introduced by Hu [1962], has been commanded as one of the most important contributions to abstract the geometric features from gray-level images. The linear transformations, such as scale and rotation changes are relative invariant in Hu's feature space. Many significant improvements based on this theory, such as [Moghaddam and Pentland 1997], [Rui 1999] and [Teague 1980], have also proved robust in applying pattern analysis tasks. Although GMI features can be extracted effectively from high quality signals, their robustness often has to be maintained by many pre-processing steps i.e. noise removal.

Features extracted from contours are also commonly used for recognition. Based on many well-developed edge detection approaches such as Sobel [Szeliski 2010], Canny [Forsyth 2003] and Prewitt [Szeliski 2010], object contours can be abstracted conveniently. For example, Wang and Xu [2010] investigated a human detection technique through human locations by combing segmentation and morphologic operations. It detects human head shapes before combining with other filtered results, such as limbs and torso based on spatial distances. Compared with the complex task

of modelling an entire human body, the shape of a human head is much simpler to define due to its relatively rigid "Ω" shape even when facing different directions, as shown in Figure 2-1.

Figure 2-1 "Ω" contours and matching result

- Statistical Features

Features defined by statistical methods have become more popular in the last 2 decades partially attributing to the growing interests in face and gesture recognitions. Extracting statistical features usually requires various transformations from 2D image space to other *n*-dimension spaces.

One classic statistic feature abstracted from gray-level images is the Haar-like feature introduced by Viola and Jones [2004]. Motivated by the early work of Papageorgiou *et al.* [1998], the Haar-like feature was initially applied in face detection applications with considerable success. This approach is sensitive to objects with high luminance contrast parts and is especially efficient in denoting objects with distinctive shape features.

- Dynamic Features

The demand for better understanding of video features and the practical call on automated video processing systems has motivated researchers to investigate dynamic

feature extraction methods and techniques. Different from the other two categories, typical dynamic features usually contain not only spatial but also temporal information that cannot be extracted from a single image or video frame. Beauchemin and Barron [1995] introduced the so-called energy flow for 2D motion analysis. The baseline of this method is coming from Horn [1987] which assumes the value of the pixel intensity is approximately constant during the studied model movement over a short period of time, which can be described as:

$$I(\mathbf{x},t) \approx I(\mathbf{x}+\delta\mathbf{x},t+\delta t), \qquad \text{2-2}$$

where $I(\mathbf{x},t)$ indicated the pixel value at location $\mathbf{x}$ and the time t. The pixel can be found at location $\mathbf{x}+\delta\mathbf{x}$ after a short period $\delta\mathbf{t}$ in a different frame, which can be applied in applications such as predictions [Baker *et al.* 2007] and tracking [Inoue *et al.* 1992].

The three categories briefly introduced above highlighted the most popular feature extraction techniques used in today's computer vision systems. Based on different applications, specific techniques from these categories will need to be carefully selected or integrated for describing different patterns accurately.

2.2. Spatio-temporal Volume Model

The fundamental data structure for defining the feature space used in this research is based on 3D volume. In simple terms, volume data can be treated as arrays of 3D vectors. The element of a volume model is the so-called voxel (an acronym for volume-pixel). One classic application of volumetric models is the 3D visualization technique for X-ray Computed Tomography (CT) and Magnetic Resonance Imaging

(MRI) scanning, which acquires 2D slices to reconstruct "solid" 3D models that can be studied and manipulated using various optical formulations.

- Defining Spatio-Temporal Volume

As illustrated in Section 2.1.2, an appropriate feature for video event detection should be constructed along on both the spatial and temporal dimensions. Comparing with conventional spatio-temporal modelling approaches that extract feature from consecutive 2D frames, the research aimed at investigating a more intuitive representation of feature space directly before applying 3D feature extraction techniques for further analysis.

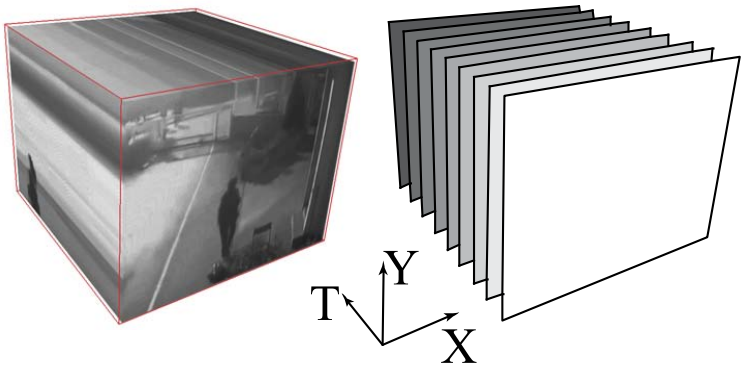


Figure 2-2 Definition of Spatio-temporal Volume

As illustrated in Figure 2-2, the STV defines a 3D volume space in a 3D coordinates system denoted by X, Y and T (time) axes. In a more observant manner, a STV model is composed of a stack of 2D arrays of pixels projecting along the orthogonal path parallel to the temporal axis. In this structure, the concept of an individual frame and its "feature" is replaced by an analogical voxel where its density, envelops and other characteristics are encapsulated in the volume space. The STV data structure enables the video event detection process to distinguish from a conventional frame-based mechanism such as optical flow becomes a real 3D analytical process. Through this transformation, dynamic information can be defined, extracted, and processed as

global features rather than the most frame-based empirical local features. Conventional 2D image pattern recognition methods, shape analysis and matching algorithms are anticipated to be developed to adapting the 3D and volumetric natures of the video events. The detailed benefits of using the STV global features for event detection will be discussed in Section 4.1.1

- Conversion between Video and STV

The main challenge of building and using volumetric event model is caused by its substantial memory consumption. For example, an uncompressed 10-second video clip of 512×512 resolution recorded at 30 Frames per Second (FPS) consumes 43MB of memory space. In contrast, some advanced video compression algorithms can dramatically reduce the size of a video file without losing too much data details.

Since the algorithms used in this project are mainly based on high computational complexity mechanisms such as sliding filter windows, iterative and looping operations, the data size of built STV models seriously affects the system performance. To ensure the system efficiency, this research only applied 3D volumetric-oriented processes on temporal related steps such as STV segmentation and event template shape matching. Additional steps on time independent feature processes such as noise removal were carried out on per-frame basis.

As illustrated in Figure 2-3, the initial STV model building operation takes the form of the so-called First-In-First-Out (FIFO) mode. The “en-queue” operation pushes frames into a queue frame by frame (FBF) following the time order, which is released (“de-queued”) when all related operations are completed. This queuing function enables a dynamic “volume buffer” structure, which is refreshed in each frame by slipping across the entire input video footages. To further improve the proposed

system, a more efficient buffering mechanism is devised in the research, which will be discussed in Section 5.3.2 with more details.

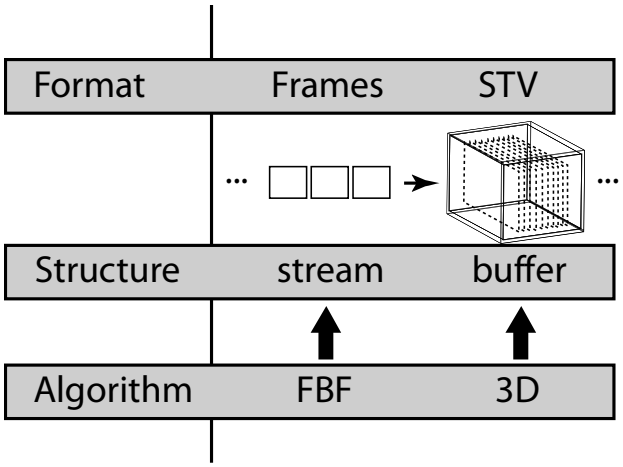


Figure 2-3 STV model construction operation

2.3. Segmentation Methodologies

The feature extraction technique used in this research is based on segmentation strategies. The segmentation in image processing and computer vision intends to produce a compact representation of the original data set by dividing through an image, a frame or a STV model into several sub-regions, which that could potentially bridge the gap between low-level features and application semantics. For example, after applying segmentation to a STV model, each voxel can be tagged by a customised label which represents the elements involved in each sub-region.

It is often confusing by looking at the segmented components alone without knowing specific applications. This can be illustrated by using a classic 2D image processing example shown in Figure 2-4. The image is a famous psychological test which contains multiple human models. In the Figure 2-4 (b) and (c), two different segmentation results introduce completely different understanding of the image with

the first one showing the silhouetted of a side-face and the second one a chatting couple. In this example, pure computer vision and segmentation techniques might not offer any "extra" features to facilitate human's perspective and cognitive processes.

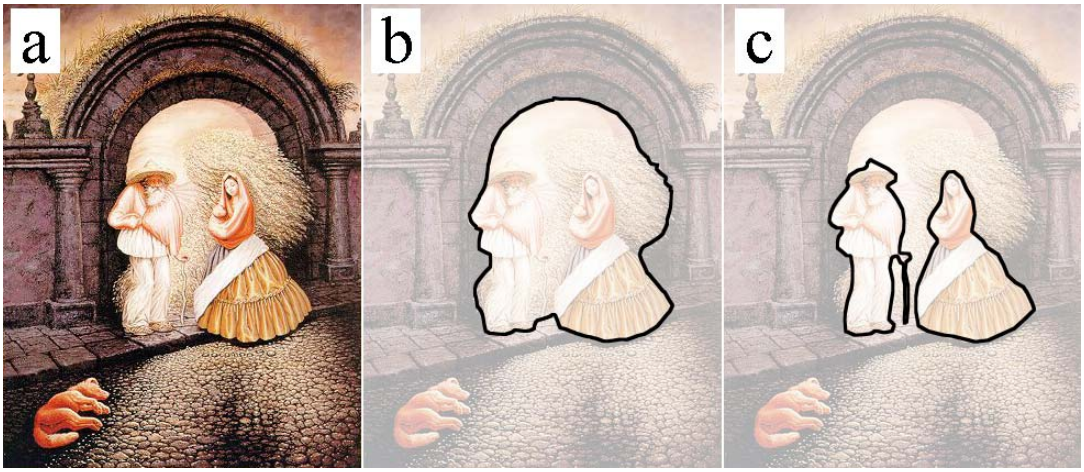


Figure 2-4 Different segmentation outputs based on same low-level features but different applications

This research will focused on the low- and intermediate-level features for building 3D geometric shapes before applying them in geometric-based pattern recognition operations.

2.3.1. Image Segmentation for Feature Extraction

Real-world scenes are filled with colours and textures. These characteristics can be easily classified by the intuitive human vision system. So far, colour-based segmentation methods employed in computer vision have seen various degrees of success in many applications with some claimed to be even better than human vision cases. But for texture, generally considered as one of the higher-level and content-based features, segmentation has proven a tough challenge.

As illustrated in Figure 2-5, the 2D artificial image can be easily divided into 3 regions visually - one shading area from white to gray, one solid colour ring, and a circle filled with high-frequency noise. In this image, the shaded area (low frequency

domain) should ideally be recognised as an entire block even if it crosses a wide brightness variation range. The enclosed circular area (high frequency domain) with black-white "pattern" should be recognised as a texture and treated as a single sub-region. However, most of the existing segmentation approaches, such as [Wu and Leahy 1993], [Weiss 1999] and [Jianbo and Malik 2000], failed to tackle this region-based problem.

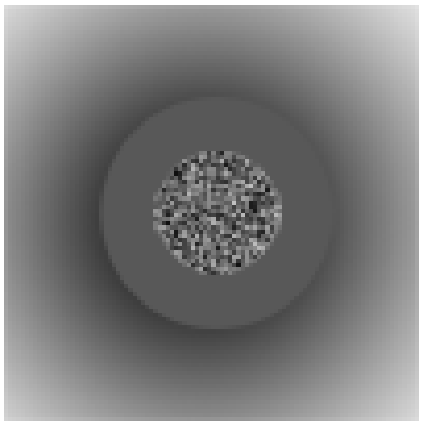


Figure 2-5 Artificial images contains gradients, solid and noise area

2.3.2. 2D Segmentation Strategies

Popular image segmentation approaches can be categorised as discontinuity- and similarity-based methods depending on the strategy applied to establish relationships between low-level features. Discontinuity methods segment images into distinctive areas by calculating boundaries between regions. These boundaries can be extracted by edge detection filters [Forsyth and Ponce 2003] or using the geometric model based Hough transformation [Duda and Hart 1972]. By contrast, similarity-based methods, such as the Region Threshold [Szeliski 2010] and the Region Growing [Adams and Bischof 1994] techniques, classify different regions through searching and organising feature points into different groups containing similar features, where features are distinctive.

The discontinuity and similarity concepts have been modelled by different algorithms that can be summarised into three types: Clustering Methods, Geometry Fitting and Probabilistic Methods

- Clustering Methods

These segmentation methods make decisions on if a component "belongs to a same group" based on pre-defined feature characteristics. These groups, known as clusters, organise feature sets into several sub-regions. The clustering methods are often unsupervised learning algorithms, which pre-define multiple "containers" for feature sets before classification. During the segmentation operation, each feature point is assigned to an appropriate container. In many cluster-based segmentation approaches such as K-mean [Forsyth and Ponce 2003], Mean Shift [Comaniciu 2002], Fuzzy C-mean [Ahmed *et al.* 2002] and Graphic-based algorithms [Forsyth and Ponce 2003], the boundaries between containers are renewed while new elements are introduced.

Clustering is a simple and flexible segmentation technique due to its unsupervised mechanism, where the feature spaces are easily composed. Therefore, clustering methods are widely used in time-sensitive applications. Another characteristic of clustering methods is induced by their flexibility on feature definition, either in "low" or "high" dimensional feature spaces. This advantage offers great benefit to this research in enabling volume-based segmentation methods from 2D to 3D. (Detailed in Section 3.2)

One common problem for clustering is due to its rigidity one feature point can only belong to one container. Points close or on the boundaries of clusters may be assigned to incorrect groups (the "under-segmentation" problem). In addition, if all existing clusters are not suitable to some specific feature points, new clusters will be generated

even if the clusters contain only one element (the "over-segmentation" problem). Solutions of these problems during event detection are devised in this research and discussed in Section 4.3 and Section 5.2.

- Geometric Model Fitting

In many applications, the contents of interest in an image or video are predefined involving geometric shapes such as lines, curves, polygons, flat surfaces and circles. Model fitting-based segmentation approaches can find feature points belonging to a particular distribution to satisfy specific geometric shapes.

The model fitting methods belong to discontinuity-based feature grouping strategy, which usually start from extracting spatial features and defining underlying geometric models. In practice, the model fitting approach is popular in some machine vision applications which can produce standard inspection. For example, in automated Printed Circuit Board (PCB) tooling systems, this technique can provide essential information on abnormality regarding sizes and locations of electronic components, welding spacing, and printing qualities.

- Probabilistic Methods

Different from the aforementioned local feature-based clustering and the model fitting methods, probabilistic methods segment image contents in global feature space. Probabilistic approaches, such as Wiener filtering-based background maintenance [Toyama *et al.* 1999], dynamic belief network [Koller *et al.*1994] and adaptive kernel region scanning [Mittal and Paragios 2004], represent the features globally by predicting some unknown parameters based on probabilistic theories, which are more robust to signal noise the local feature-based approaches.

2.4. Pattern Recognition Approaches

After extracting available features from videos, the challenge for event detection moves onto the selection of appropriate pattern analysis methods. Based on the specific techniques applied during recognition, pattern analysis methods can be categorised into two main types: template matching and machine learning.

- Template Matching

A template is often referred as pre-defined representative models containing chosen features as illustrated in Figure 2-6. The template matching mechanism is a comparison process performed in the feature space by measuring the "distance" between at template and the unknown patterns based on their feature distributions. In addition, the dashed line showing in the figure denotes the system maintaining steps, which improves the image features represented by the templates after evaluating the performance of the system accuracy.

For example, in 2009, Cui [2009] developed a template matching algorithm for matching 2D open curves. In this algorithm, the distance is calculated based on the cross-correlated confident normalised covariance [Lewis 1995]. The basic theory of the method is to compare two curves by correlating and evaluating curvature similarities through employing a curvature integral, which significantly reduces the problem caused by scaling and rotational transformations.

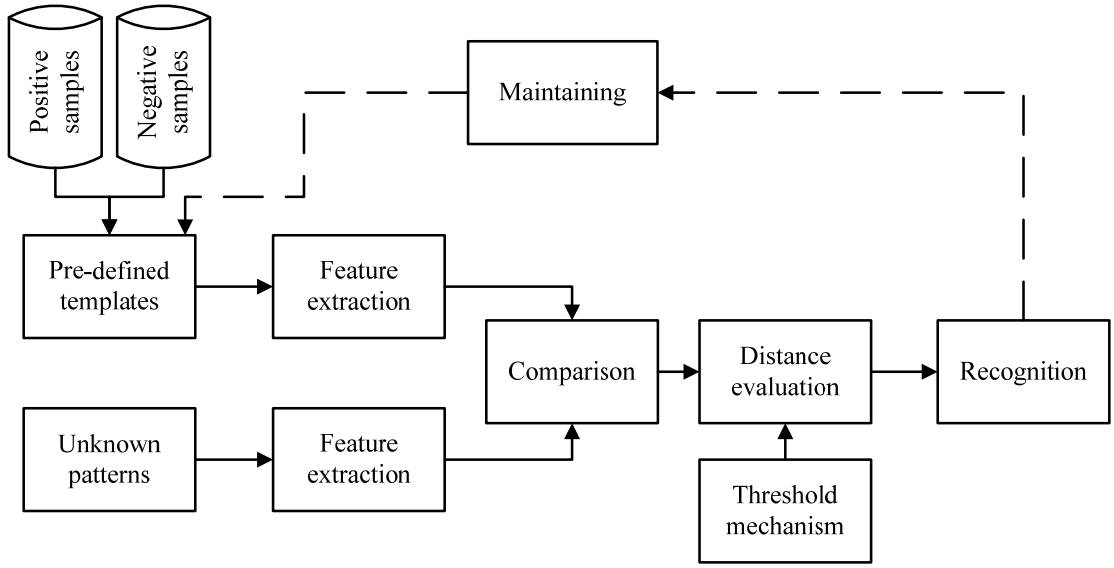


Figure 2-6 Template matching system pipeline

- Machine Learning

Compared with template matching, the machine learning approaches do not require model-matching during the recognition operation. The advantage of this method is that the system robustness can be strengthened and accuracy improved after each recognition cycle through feed-in adjustments generated from current detection results (as illustrated in the dashed line denoted in Figure 2-7).

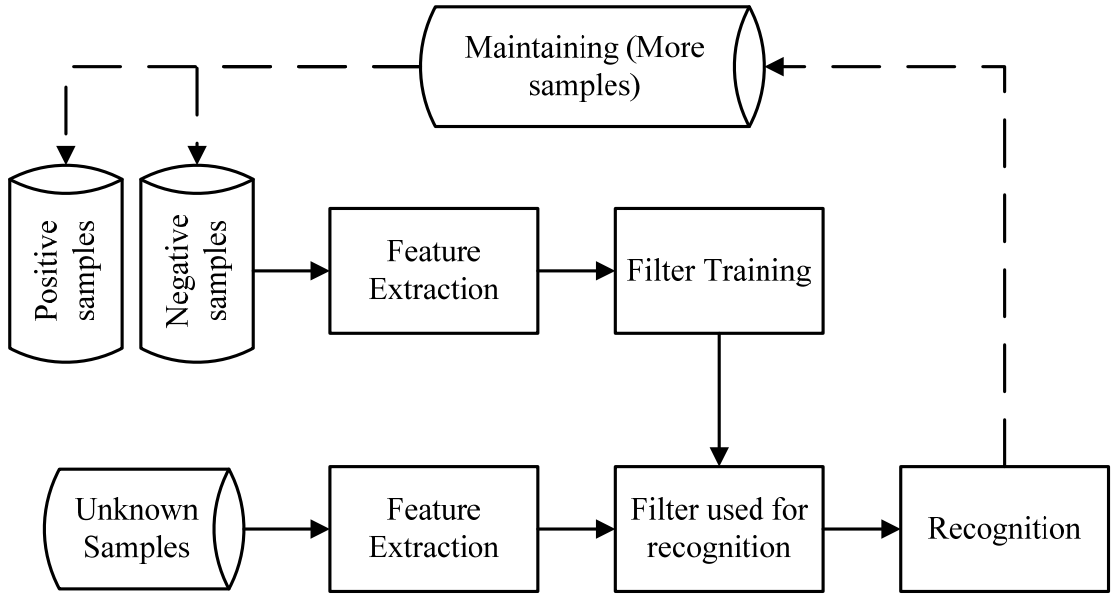


Figure 2-7 Machine learning system pipeline

But this model-free approach is at the expense of many preparation steps known as learning process. Based on different implementation strategies, the learning processes are used for composing feature classifiers [Scott 1992], decision trees [Quinlan 1986] and artificial neural networks [Hopfield 1988] for recognising unknown patterns.

2.5. Video Event Definition and Human Action Detection

Video event detection is a popular research area in computer vision aiming at finding and understanding pre-defined real-world "events" in a video in an online or offline style. In this research, a narrower definition of video events has been adopted that focuses on single human actions and gestures, which often enable applications in surveillance and security areas. The common process pipeline of video event detection integrates local or global event feature extraction, classification and recognition. Recent surveys on video event detection techniques by Moeslund [2001; 2006] highlighted many applicable detection techniques in details, for example, the parking surveillance system, motion capturing for human interaction and performance analysis for athletes skill improvement.

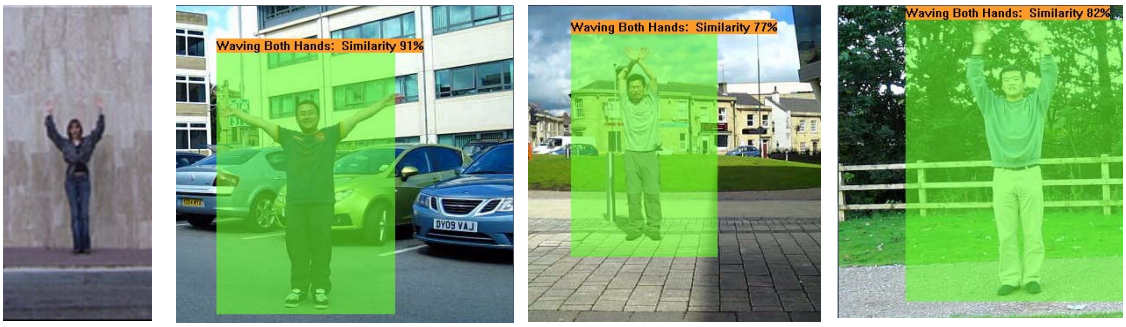


Figure 2-8 **(a) "Waving" template, (b), (c), (d) selected snapshots of detected events**

In this research, typical human gestures such as waving, walking, jumping, and running can be dissected and encapsulated into action "tablets", or so-called "atomic events" [Reng *et al*. 2005] carrying semantic values and being further processed, for example through segmentation and recognition.

The event detection system developed in this research had been focused on automatically detects and identifies real-world human activities. As illustrated in Figure 2-8, (a) is a "waving hand" event extracted from a public human activities database (see Section 6.1). The time duration of the event tablets set at 2 seconds [Gorelick 2007]. Figure 2-8 (b) to (d) are the snapshots of 3 identified "waving" events used as pattern videos for recognition which are approximate to 2 seconds each. It is obvious in this example that the background of the action template is relatively "clear" but the untreated videos were filled with background noise, i.e. texture patches, passing vehicles, and pedestrians.

In addition, application-specific postures and posture changes can also be defined as "events". For example, Figure 2-9 shows a "falling down" event extracted from the real-world CCTV recordings downloaded from the YouTube website. The video shows many pedestrians fell over on a pedestrian path near the entrance of a building. This application also demonstrated the need to track "changes" over video frames that are often affected by poor quality on video signals, in terms of low resolution, illumination variations, occlusion problems and "semantic" ambiguity.

Figure 2-9 Application-specific "fall down" event

The main theoretical approach of this research is to integrate spatial and temporal features of the studied subject within a unified global space, where 3D volumetric features, taking the form of groups of "related" voxels can be analysed. The shapes and densities of those voxel "clouds" can represent object movements, silhouette transformations and even reveal "internal" characteristics such as "gain" patterns. One of the anticipated drawbacks of the volume-based approaches is its intrinsic difficulties in accessing and processing data in real-time. In this research, the algorithm-based optimisation and acceleration methods have been investigated with their potentials on improving STV-based operations discussed. There are two main focuses in this project: the investigation and development of effective volumetric event detection techniques based on innovative template matching algorithms; and the evaluation of the corresponding STV processing technique to enable "on-fly" volumetric event models/templates construction.

2.5.1. Current Video Event Detection Research and Practices

Generally speaking, a video event can be classified as single-human-based, such as gestures and postures; multi-human-based like crowd behaviours; and non-human-based, for example, vehicle or abnormal object movements. For specific research

problems or applications, many pilot works have defined specific conditions to simplify the settings of system platforms. A survey on those conditions can be found in Popper's Survey [2010]. The downside of this approach is the rigidity of the algorithms and its compromised suitability for challenging real-world settings.

Single-human-based events have often been used for tracking individuals, recognising gestures, and monitoring the change of behaviours. [Guler *et al.* 2007] has published a paper on tracking individuals who had left a baggage behind and developed a real-time prototype system for verifying the concept. A number of research papers have summarised progresses in this area such as [Zhou and Hu 2008] and [Matikainen et al. 2009]. Popular computer vision databases such as KTH [Schuldt *et al.* 2004], Weizmann [Gorelick 2007] and Inria XMAS [Weinland *et al.* 2006] have also been focusing on adopting single-human models including shape, contour and stick skeleton for analytical tasks. [Wang and Suter 2007] also developed a human action recognition system through analysing human silhouette and Locality Preserving Projections (LPP), which reduced the dimensionality required for transforming human actions into low-dimensional spatio-temporal feature space. This development had improved the system tolerance on problems such as partial occlusion and noisy background.

For multi-human-based events, particle flows and density models are often deployed to analyse crowd behaviours with individuals being treated as moving particles. Kilambi *et al.* [2006] have adopted a Kalman filter-based approach for estimating human group sizes which demonstrated superior performance over the traditional shape or Bayesian-based method. In addition, dynamic optical flow system [Fleet *et al.* 2000] and trajectory matching algorithms [Porikli 2004] have both been deployed

based on the Hidden Markov Model (HMM) for handling the random movements of the monitored crowds. Research progresses on this front have seen concepts and prototypes developed for applications such as intelligent and adaptive traffic light control, and crowd emergency evacuation systems. In contrast to the single-human-based events, crowd behaviours and events are sometimes difficult to define using explicit with accurate semantic interpretations.

A wide range of non-human-based events exist that focused on the changes of arbitrary objects' shapes, locations and other physical "statuses". Compared with human-based events, these events often carry more image or pixel-level features. For example, an explosion can be described by the rate of colour, illumination, and shape changes for groups of pixels. Liu and Ahuja [2004] devised a fire detection system based on this principle, which employed spectral models containing colour templates describing contours of fire and the Fourier features at the frequency spectrum. Real-time vision-based traffic control systems [Coifman *et al.* 1998] and smoke detection system [Wieser and Brupbacher 2001] have also defined non-human events with clear semantic meanings.

The research in this project has been concentrated on improving single-human-based event detection technologies with special attention focused on practical spatio-temporal model for conceptual and contextual-level understanding of video-based events.

2.5.2. Reviewing of STV-based Event Detection

Stemmed from DIP techniques, traditional video event detection approaches relied on the spatial or frequency features being extracted from the per-frame-based 2D processes [Gavrila 1999]. However the Frame-by-Frame (FBF) mechanisms often

resulted in the loss of "contextual" information – a main contributor to defining "dynamic" video events. This drawback can lead to high false-positive rate during an event detection operation, for example, mixing an action or pose from two different persons crossed over in front of a video camera. One perceived solution for this problem is to construct video capsules which encapsulates both time-tag and frame patterns.

A number of leading computer vision research groups have devised various integrated spatial and temporal techniques for video analysis, such as flow-based iterations [Beauchemin and Barron 1995], motion history image [Bobick and Davis 2001], and local interest points [Shi and Tomasi 1994], which have focused on the "durations" and "changes" of spatial features over time. But most of these features are gathered from consecutive frames on a small group of pre-determined pixels. The global changes over the entire scene and the 3D event exercisers – humans or objects – cannot be represented in their entirety.

Constrained by the performance of computer capacity at time, most of the preliminary research on this innovative 3D voxel-based data processing and pattern analytical technique largely remained at the level of theoretical discussions until the middle of 1990s, when a number of world-leading image processing research groups had attempted to map the STV models to their customised 2D projection planes before applying the normal pixel-based processing methods for further analysis. One of the representative methods from those approaches was using the so-called "clipping plan" for feature extraction along the time axis on a volume. The slice-based STV models have been successfully adopted to infer feature depth information [Baker and Bolles 1989], generating dense displacement fields [Li *et al.* 2001], analysing camera

calibration settings [Kuhne *et al.* 2001], categorising human motion patterns [Ngo *et al.* 2003], and performing viewpoint synthesis [Rav-Acha and Peleg 2004].

For example, a human gait analysis method based on STV slices was introduced by [Niyogi and Adelson 1994]. As illustrated in the Figure 2-10, the "Y"-value of the slice was chosen at knee-height in the volume. The output "image" is composed by clipping the STV parallel to the X-T plain. The 2D image contains braid-style textures, where the feature of the gait can be abstracted effectively. For example, each joint of the braid represents the location and moment of the feet alternation; the slope defined by the line across these joins denotes the speed of the walking. Other information such as step length and cycle can also be easily extracted from this XT slice.

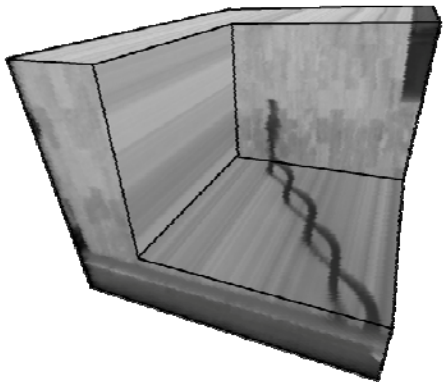


Figure 2-10 STV slices used for human gait analysis

Since the start of the new millennium, the "real" volumetric approaches for STV processing have been steadily gaining popularity. These new approaches have taken advantages of the 3D volumetric natures of feature points and emphasises on the alteration of shapes, envelopes, and density of those points in an enclosed space. Research advancements in many related areas such as volume visualization [Yeung and Boon-Lock 1997] and medical image processing [Peng *et al.* 2010] have contributed to the development of these improved voxel-based techniques. Generally speaking, most volumetric approaches have focused on global representations of the

studied subjects and are following a "top-down" [Popper 2010] processing pipeline which contains several phases such as global tracking, segmentation, modelling and recognition. Global representation has been proven as a valid and "cost effective" approach [Due *et al.* 1996] for most of the video event detection tasks. More recently, other advanced approaches that integrate the selected local features, such as optical flow, with STV-based global features have been proposed for specific systems and applications. A state-of-the-art review on STV and its applications carried out in this project has identified some important works in the field, for example, spatio-temporal cuboids prototyping methods [Dollár *et al*. r2005], inter-frame constrained local feature definitions [Jiang *et al.* 2006], and STV Bag-of-Words-based (BoW) volume feature points [Siva and Xiang 2010]. These approaches can be divided into geometric shape-oriented and spatio-temporal interest point-oriented method.

- Global Feature-based Geometric Shape-Oriented Event Definition

Shape-based STV methods treat original STV data sets as "sculpture"-like geometries formed by the distributed "point clouds". Alper and Mubarak [2005] introduced a method to extract 3D human silhouettes from the volumetric space for shape matching. Based on the shape invariants, Alper's method had applied the so-called "differential geometric surface properties", such as peaks, pits, valleys and ridges as feature descriptors to denote specific events in the form of vectors in a feature space.

In 2006, through "transplanting" motion history images onto the 3D STV model, [Weinland 2006] developed a set of view-invariant motion descriptors for human event definition that is capable of representing dynamic events captured in a video by applying Fourier transformations in a cylindrical coordinates system. This process

formed a solid foundation for the extraction of view-invariant features generated from event templates in the form of patterns in the frequency domain.

A 3D shape-invariant analysis method, which was first proposed by Gorelick *et al.* [2007] is another significant progress in the field. By deploying Poisson distance equation $U_{xx}+U_{yy}+U_{tt}=-1$ and its Dirichlet boundary condition $U(x,y,t)=0$ [Gorelick *et al.* 2006] inside the STV-segmented shapes, the local space-time saliency features, $U+1.5\|\nabla U\|^2$, and the Hessian-based space-time orientations features (the ratio of the eigenvalues of Hessian matrix of the Poisson equation), generated to describe different volumetric shape features, which are sensitive to the unique STV shapes generated by arms, torsos and legs. Based on the method, human gestures can be categorised into a number of types by applying a spectral classification algorithm [Tangelder and Veltkamp 2008].

- Local feature-based Spatio-temporal Interest Point-based Methods

Spatio-temporal interest point was first introduced to the STV domain in 2003 by [Laptev and Lindeberg 2003]. Using this technique, the interest points can be effectively located on the rapid changing sections in a video sequence. By representing these "changes" using the so-called "Bag-of-Words" [Li and Perona 2005] - a novel pattern recognition method - different events can be classified into various categories. [Niebles et al. 2008] had attempted in addressing the action categorization problems by using the "spatio-temporal words" - an extended version of the original "Bag-of-Words" method, which was facilitated by an unsupervised learning algorithm.

[Willems *et al.* 2008] devised a more robust model based on the same concept to tackle the dense- and scale-invariant interest point problems. By using the Hessian as a saliency measurement, feature points can then be represented invariantly both in the spatial and the temporal domain. Willems also proposed and tested the so-called "box-filter", a 3D version of Haar-like filter, as a feature detector. Compared with other popular methods such as [Laptev and Lindeberg 2003] and [Oikonomopoulos *et al.* 2005], Willems' method had shown superiority in terms of repeatability, accuracy and speed.

2.6. Application Domains

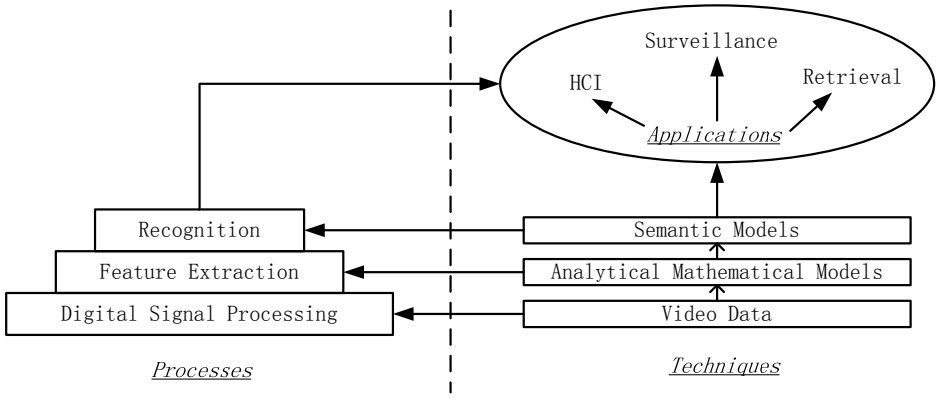


Figure 2-11 System hierarchical structure

Figure 2-11 illustrates the process paradigm of a classic video event detection system. Based on fundamental studies over the last 30 years on DIP, feature extraction, pattern analysis and biological vision, variation modern video event detection techniques have been applied in a number of domains. Compared to traditional human-in-loop operational mode, the automated processing paradigm is "enjoying" a degree of success while still facing many challenges. Table 2-1 compared 3 main

application domains where event detection techniques have been widely expected to make a substantial impact.

Application	Real-time	Accuracy	Controlled environment	High performance hardware
Human Computer Interaction	★★★★★	★★★☆☆	★★★★☆	★★★★★
Surveillance	★★☆☆☆	★★★★★	★☆☆☆☆	★☆☆☆☆
Video Search & Retrieval	★☆☆☆☆	★★☆☆☆	★★☆☆☆	★★☆☆☆

Table 2-1 Application domains and performance demands (more solid stars means higher demand)

In HCI, various computer games are pioneering the more intuitive and "intimate" interaction styles with computers via vision and other sensory devices, for example, Microsoft Kinect [Suma *et al.* 2011] and Six Sense projector [Zoran and Coelho 2011]. These interaction devices track human gestures or posture changes in real-time based on advanced hardware support and innovative vision software algorithms. These systems are often established in controlled environments with steady indoor lighting conditions and small detection ranges.

Surveillance is another important application area for the video event detection. For example, to detect various vandalism actions, such as graffiti drawing, the event detection system can assist an "early warning" mechanism through tracking certain pre-defined suspicious behaviors "over a period of time" before triggering or disarming alarms. However, compared to HCI applications, the event detection based intelligent surveillance systems are often based on more complicated "decision-making" processes, hence, more suitable to an off-line operational mode. Other problems facing the surveillance and security applications include low-resolution visual signal, varied illumination conditions and longer-range detection scope.

The rapidly increasing digital media repositories and the explosion of web-based image and video databases present another challenge (and opportunity) to computer vision research in effective and efficient digital library management, for example, in sporting video analysis and editing, the problems such as keywords ambiguous and translation between different languages can be tackled by using CV-based video management techniques rather than traditional text-based retrieval system.

2.7. Prominent Challenges for Video Event Detection

The complexity exposed to video event detection tasks can be classified into three categories. Firstly, the boundary between an "event" signal and its "background" noise is often inexplicit, which renders the separation of the two signals impossible. In most of the current approaches, the background is often simplified as static sections in continuous video frames. However, this presumption is not always applicable in a complex scenario with multiple moving objects existed. Secondly, the semantic of an "event" in a video is ambiguous since the variations of potential "event makers" that are often defined by a particular application. For example, illumination, colour, shape, or texture changes over a defined period of time. Another difficulty is caused by the uncertainty of durations of video events. The time-elapsed factor for encapsulating a discrete event is closely coupled with video sampling rates which might be substantially varied for different videoing hardware.

Finding solutions for these problems in specific cases are also the objectives of this research. In the following part of the thesis, these solutions will be discussed in detail in each chapter.

Chapter 3. Segmentation-based Shape Feature Extraction

As mentioned in Section 2.2, STV voxels contain many low-level features such as colours, intensity and location, which provide fundamental data for constructing semantically higher-level features for different applications. In this research, the abstraction is mainly based on standard 2D and extended 3D segmentation techniques. The segmentation results intend to provide robust and sophisticated volumetric shapes and regional features for the following pattern recognition steps.

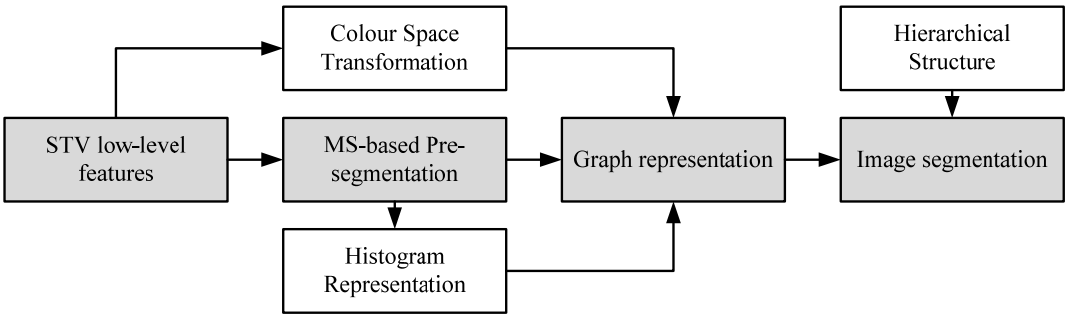


Figure 3-1 Segmentation pipeline used in this research

In this research, an innovative segmentation algorithm is developed based on a hybrid discontinuity and similarity segmentation model by combining Mean Shift (MS) clustering and the graph-based region description method. The progress pipeline of this approach is illustrated in Figure 3-1, which starts from a pre-segmentation process by using MS clustering. This straightforward and rapid operation provides roughly segmented sub-regions for graph-driven refinement using the STV sub-regions rather than the primitive voxel data. The baseline graph representation algorithm is stemmed from the earlier research introduced by [Feizenszwalb and Huttenlocher 2004] and [Grundmann *et al.* 2010]. This advanced algorithm organises

the sub-regions based on their similar appearance in a graph. As shown in Figure 3-2, the artificial image in Figure 2-5 can be precisely segmented into 3 sub-regions that are to resemble human vision and perception functions.

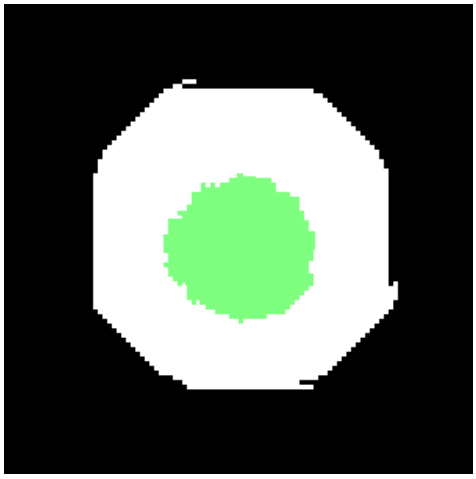


Figure 3-2 Segmentation output of Figure 2-5 by using newly developed approach introduced in this research

In this research, improved clustering methods are used for video pre-. The rest of this chapter is organised in the following order: The baseline 2D graph-based segmentation algorithm is introduced in Section 3.1, Section 3.2 highlights the investigation and development of an innovative 3D volumetric segmentation approach. In section 3.3, the outputs of the devised algorithm is tested and evaluated by real-world video footage. Section 3.4 concludes the work in the feature extraction stage and its relations to the following event detection steps.

3.1. Baseline Methods for STV Feature Extraction

One of the most important features used in this research is 3D shapes extracted from the STV. For representing shape features geometric distribution in a STV model, an Improved Pair-wise Region Comparison (I-PWRC) clustering segmentation method has been devised in this research. As introduced in Section 2.3, the non-supervised

clustering methods are considered more efficient than the geometric and probabilistic methods since the calculations on large amounts of STV voxels and maintenance checks on the geometric parameters can be extremely time-consuming.

The baseline segmentation method adopted in this research is call "Pair-wise Region Comparison" (PWRC) [Feizenszwalb and Huttenlocher 2004]. This graph-based clustering method is a capable segmentation approach for classifying similar textures based on original intensity or colour features. The algorithm is based on a typical iteration mechanism and renews each cluster by comparing the so-called inner difference between neighbouring elements. In this research, this 2D graph similarity comparison mechanism is extended into 3D feature space for facilitating the video event detection demands in STV feature space.

3.1.1. Baseline 2D PWRC

This segmentation approach organises sub-regions based on their similar appearances in a graph $G=(V,E)$ where V denotes a collection of vertices v_i in the graph and E denotes the collection of edges e_i between two vertices that $(v_i, v_j) \in E, (i \neq j)$. In the graph, edges are used to denote the differences between two vertices by assigning a weight value $w[(v_i, v_j)]$, (i≠j) to each edge. When applied, the PWRC evaluates image features by looking at the differences between two weighted edges. It is worth noting that the "difference" can be measured in spaces often represented in a multi-dimensional vector format composed of elements.

Based on the Graph theory and Felzenszwalb's and Huttenlocher's work, the image segmentation operation S is a clustering process of V. Components of S includes many subsets in the graph. In an ideal situation, the segmentation outputs should contain several segmented clusters C. The features involved in C_i, (i=1,2,…,n) should be

identical, and features in different clusters should be distinctive and separable. In addition, the weights of edge subset in C_i, should be relatively small compared to edges of any two vertices from different clusters.

The output of the algorithm is a predication value, D, to determine if a boundary exists between regions, with each vertex needs to be located in an independent cluster C_i, (i=1,2,…,n). The similarity between different regions and the dissimilarity inside a region can then be compared. Depending on the thresholding "similarity factor", two regions can be merged into a larger region. During the operation, the initially independent regions will keep evolving in an iterative style until reaching a "balanced" stage defined by the threshold.

To model this process, a so-called "internal difference" $Int(C)$ is defined for evaluating the element differences in a cluster, which is looking for a maximum edge weight in its Minimum Spanning Tree (MST) as shown in Equation 3-1.

$$Int(C) = \max_{(v_i, v_j) \in MST(C,E)} w((v_i, v_j)), \quad \textbf{3-1}$$

where w denotes the weight of an edge. $Int(C)$=0 if C contains only one vertex element. Since the MST defines a minimum cost description in a graph, other components in the same connected graph should contain at least one edge that is larger than $Int(C)$, which indicates the lower threshold of the internal feature difference.

Equation 3-2 represents the minimum difference between two distinctive clusters which is the lower threshold for merging two regions. The $Diff$=∞ if there is no edge connecting C_1 and C_2.

$$Diff(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} w((v_i, v_j)), \quad \textbf{3-2}$$

The predication then follows based on the calculated differences between the two clusters and their internal difference, that is:

$$D(C_1.C_2) = \begin{cases} true & Diff(C_1.C_2) > MInt(C_1.C_2) \\ false & otherwise \end{cases}, \quad \textbf{3-3}$$

and

$$MInt(C_1, C_2) = \min\left[\left(Int(C_1) + \frac{k}{|C_1|}\right), \left(Int(C_2) + \frac{k}{|C_2|}\right)\right], \quad \textbf{3-4}$$

where $|C|$ denotes the cluster size and k stands for a constant parameter to control the "sensitive factor" of segmentation.

At the start of this algorithm, the entire graph is composed of only pixel values in an image. By calculating $Diff(C_i, C_j)$ and $MInt(C_i, C_j)$, regions can be formed by merging 2 clusters if the difference between them is even smaller than their own inner differences. Since the regions are self-independent in the undirected graph, the region comparison can be initialised from any cluster. After traversing all edges in the graph, D represents a stable state with several regions formed; separating different feature points and storing similar ones in each region. In this algorithm, when $|C|=1$, since the internal difference is equal to 0 and $Diff(C_i, C_j) \geq 0$, the internal merging operation cannot be performed or to compare $Int(C)$ with $Diff(C_i, C_j)$ directly. This problem is resolved by adding a compensation factor $k/|C|$ as illustrated in Equation 3-4.

The sensitivity of the PWRC segmentation is controlled by the coefficient k with larger values leading to bigger segmented regions. When smaller values are applied, it can ensure most of the important boundaries are extracted. This research has chosen

smaller k for extracting 3D shapes as inputs for the following matching operations. Figure 3-3 provides snapshots of sample PWRC segmentation results.

Figure 3-3 Snapshots of PWRC segmentation outputs

3.1.2. I-PWRC for STV Segmentation

As explained in Section 2.1.1 and 2.3, many segmentation approaches can be extended into 3D domain, including PWRC for video data segmentation. For STV, the I-PWRC graph is initialised by 3D features. Vertices are represented by volumetric features with the 26 connected neighbours of each voxel constructing the edges of the graph. Since the edge contains both spatial and temporal information, the alterations and "traces" can reflect the dynamic stages of the "tracked" objects during segmentation compared with the "static" feature-only FBF-based techniques.

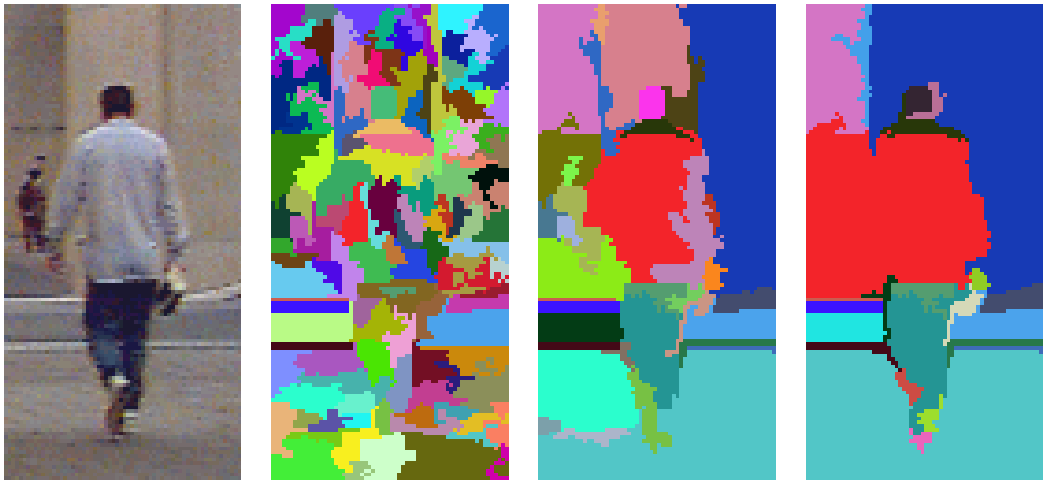


Figure 3-4 Factor *k* is not flexible to deal with complex real-world environment

During the initial feasibility experiments, it was observed that direct transformation of the 2D to the 3D PWRC led to several drawbacks. Firstly, the region *C* cannot be readily controlled by the factor *k* comprehensively. As illustrated in Figure 3-4, larger *k* values lead to larger segmented regions but missed out some segments of object boundaries. Smaller *k* values ensure all the object boundaries being abstracted but in a so-called over-segmentation style with large amount over-cut and cluttered areas. Secondly, during the region growing stage of PWRC, the internal difference *Int*(*C*) becomes less sensitive especially in the regions filled with random textures since the weights of the graph is based on simple voxel values (low level features) rather than region texture features. Thirdly, to generate a PWRC graph based on raw STV introduces huge amount of data, for example, 100×100×100 STV adjacency of the size $(100\times100\times100)^2$, which cannot be readily handled by conventional computer memory, never mention other looping and branching operations on this data.

The I-PWRC, on the other hand, deployed a different STV data manipulation strategy by combining the MS segmentation with a hierarchical data storage structure. It optimised the performance of factor *k* with a region feature-based representation

scheme for weight definitions and enabled the reduction of the data sizes for more effective STV segmentation.

3.2. I-PWRC Implementation

3.2.1. Pre-clustering Using Mean Shift Algorithm

To apply PWRC to STV segmentation, the vertices in the initialised graph are un-treated voxels. However, a large percentage of those voxels only contain background information, which has no contribution to the feature template matching operations and are filtered out using the innovative I-PWRC technique developed in this research. I-PWRC simplifies the initialised graph using MS-based clustering through removing redundant low-level features and combining similar voxels into small regions represented by graph vertices. Compared with the per-voxel-based initialisation, the vertices numbers are reduced significantly.

In the I-PWRC, the pre-clustering is carried out using the MS model-seeking algorithm. Although sensitive to noise, the MS method can handle small groups effectively in feature space and controlling the segmented region sizes flexibly.

Applying the so-called Parzen window density estimator [Duda and Hart 1973] in this research, the MS clustering occurs prior to the "primary" STV graph segmentation by using a Probability Density Function (PDF).

Based on Comaniciu's paper [2002], in a d-dimensional ($\mathbf{R}^d$) feature space, if $\mathbf{X}$ denotes the collection of feature points with individual point represented as $\mathbf{x}_i$, i=1,2,…,n, the multivariate kernel density estimator with kernel $K_H(\mathbf{x})$ can be computed at the point $\mathbf{x}$ by

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(\mathbf{x} - \mathbf{x}_i\right), \quad \textbf{3-5}$$

where H is a symmetric positive $d \times d$ bandwidth matrix. and

$$K(\mathbf{x}) = c_{k,d}\, k(\|\mathbf{x}\|^2), \quad \textbf{3-6}$$

where $C_{k,d}$ is a non-negative normalisation factor and the profile k is defined by a Gate Function (GF)

$$k(x) = \begin{cases} 1 - x & 0 \le x \le 1 \\ 0 & x > 1 \end{cases}, \quad \textbf{3-7}$$

It also predigests H from fully parameterized matrix to an identity matrix $H=h^2I$, where h is the window size of the MS. It is the only bandwidth parameter need to be provided before MS operation, which is particularly suited to this project for its simplicity.

After introducing Equation 3-6 into the kernel density estimator (Equation 3-5), the proximate expression can be rewritten as

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right), \quad \textbf{3-8}$$

with its quality measurable using the mean of the squared errors between the densities and their sum over the domain.

The MS at runtime operation finds the peak values in the feature space and then classifies relevant feature points in vicinities. In the density estimator model introduced above, different peak values can belong to different maximum density areas, which mean $\nabla f(\mathbf{x})=0$. This can be explained in the following expression:

$$\nabla \hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_i) k'\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right), \quad \textbf{3-9}$$

and

$$g(\mathbf{x}) = -k'(\mathbf{x}), \quad \textbf{3-10}$$

where *k'* is the derivative of the profile *k*. Therefore, Equation 3-9 is transform into

$$\begin{aligned} \nabla \hat{f}_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}) g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \right] m_{h,G}(\mathbf{x}). \end{aligned} \quad \textbf{3-11}$$

As proven by [Comaniciu 2002], the Mean Shift vector $\mathbf{m}_{h,G}(\mathbf{x})$ in the above equation can be expressed as:

$$m_{h,G}(\mathbf{x}) = \frac{h^2 c}{2} \frac{\nabla \hat{f}_{h,K}(\mathbf{x})}{\hat{f}_{h,G}(\mathbf{x})}, \quad \textbf{3-12}$$

in the feature space.

It is proven in the I-PWRC feasibility study, by carefully selecting the window size factor *h* on colour h_c and location h_l, the regions composed by MS are of a satisfactory standard with key feature identified which can be well controlled. For example, a small window size (h_c=5 and h_l=5) was tested on a 20-second video clip for the pre-clustering process. It was observed that most important details relating to the event contents and volumetric boundaries were identified, which provided a valid foundation for the following I-PWRC and template matching processes.

3.2.2. Histogram-based region description

After the pre-clustering process, the initialisation of the segmentation graph is carried out on region features rather than voxel features that reduce the quantity of the required vertices. To tackle the feature insensitivity problem during region growing stage in the traditional PWRC (see Section 3.1.1), this research introduced tools such as histogram and histogram distances to represent region textures in the weighted graph.

- Colour representation

The research has chosen the L*a*b* colour space to define complex high level features in a histogram due to its superior adaptability to human vision cases. (See [Forsyth 2003] and [MacEvoy 2010] for more details).

There is no direct mapping and conversion algorithms from the classic RGB to L*a*b* colour space. Hence, an intermediate XYZ colour space is created by following transformations:

$$\begin{bmatrix} X & Y & Z \end{bmatrix} = \begin{bmatrix} R' & G' & B' \end{bmatrix} \mathbf{M}, \qquad \textbf{3-13}$$

where $\mathbf{M}$ is the transformation constant. Based on CIE Standard Illuminate D65 [Schanda 2007], the $\mathbf{M}$ is defined by

$$\mathbf{M} = \begin{bmatrix} 0.412424 & 0.212656 & 0.019334 \\ 0.357579 & 0.715158 & 0.119193 \\ 0.180464 & 0.0721856 & 0.950444 \end{bmatrix}, \qquad \textbf{3-14}$$

Let W denote any element in [R G B] vector, W' is unified W value based on

$$W = \begin{cases} W'/12 & W' \le 0.04045 \\ \left(\frac{W'+0.055}{1.055}\right)^{2.4} & otherwise \end{cases}, \quad \textbf{3-15}$$

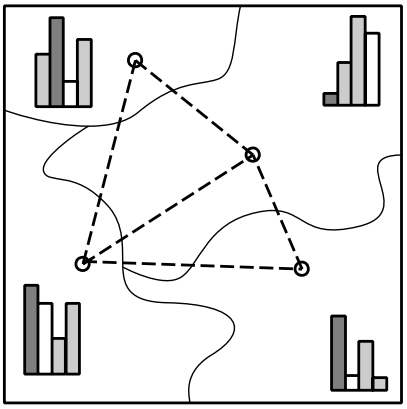
The L*a*b* transformation from XYZ colour space can then be defined as:

$$\begin{aligned} L^* &= 116 f(Y/Yr) - 16 \\ a^* &= 500\left[f(X/Xr) - f(Y/Yr)\right], \\ b^* &= 200\left[f(Y/Yr) - f(Z/Zr)\right] \end{aligned} \quad \textbf{3-16}$$

where [*Xr Yr Zr*] is a reference white colour and defined as [0.95047 1 1.08883] based on CIE Standard Illuminate D65, and

$$f(x) = \begin{cases} x^{1/3} & x > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 x + \frac{4}{29} & otherwise \end{cases}, \quad \textbf{3-17}$$

- Graph representation



After the colour transformation, high-level region features, such as textures, can be readily defined by the local colour histograms to denote vertex values of a graph. As illustrated in Figure 3-5, each region contains a normalised local histogram based on the L*a*b* colours. The distribution of these colours in the histogram embodied richer information than voxel-level features. For example, a texture containing flat

distribution of solid colours can be readily represented as multiple peaks in the histogram.

Weights of edges in this new form of graph are defined by the histogram distance. In this research, a minimum distance method introduced by Cha [2002] has been adopted, which can be abstracted as follows:

Suppose H_X and H_Y are denoted as two different unified histograms that contain n elements each with individual element specified in the style of $h_{X,i}$ and $h_{Y,i}$, where i=1,2,…n. The distance D(H_X, H_Y) of these two histograms can then been summarised as:

$$D(H_X, H_Y) = \min_{X,Y}\left(\sum_{i,j=1}^{n} d_{\text{mon}}(h_{X,i}, h_{Y,i}), \sum_{i,j=1}^{n} d_{\text{ord}}(h_{X,i}, h_{Y,i}), \sum_{i,j=1}^{n} d_{\text{mod}}(h_{X,i}, h_{Y,i}) \right). \quad \textbf{3-18}$$

In the Equation, d_{mon} denotes "nominal measurement" highlighting the existence of the histogram bins; d_{ord} denotes "ordinal measurement" for calculating the difference between the weights of each bin; d_{mod} denotes "modulo measurement" which evaluate the arithmetic modulo along the angular values such as the distribution of colours in L*a*b* colour space. Each measurement can be modelled as:

$$d_{\text{mon}}(x,y) = \begin{cases} 0 & x = y \\ 1 & otherwise \end{cases}$$

$$d_{\text{ord}}(x,y) = |x-y| \quad , \quad \textbf{3-19}$$

$$d_{\text{mod}}(x,y) = \begin{cases} |x-y| & |x-y| \le \dfrac{n}{2} \\ n-|x-y| & otherwise \end{cases}$$

Therefore, the weight defined by the minimum distance can be expressed as:

$$w[(v_i, v_j)], (i \neq j) = D(Hv_i, Hv_j), \quad \textbf{3-20}$$

3.2.3. Hierarchical Pair-wise Region Comparison

As discussed in Section 3.2, region sensitivity of conventional PWRC is only controlled by factor *k*, and a fixed *k* value for the region growing operation is insufficient for feature grouping if texture number varies substantially. This drawback has resolved in this researched I-PWRC method by introducing a hierarchical segmentation structure using adaptable and dynamic *k* values.

It is widely recognised that real-world images or video frames especially the outdoor scenes, often contain many large and uniform colour regions (i.e. sky and soil) as well as varied textures (i.e. flowers and grass). Most existing segmentation strategies are seemingly specialised in dealing with either prior or latter case. Hierarchical structure offers a practical approach to solve this problem by building a pyramid structure for storing and representing raw data, as illustrated in Figure 3-6. The low resolution images or frames at the top of the pyramid only need to "remember" large coloured blocks with minute details filtered out during the re-sampling operation. The bottom level, on the other hand, records all the details of the original dataset.

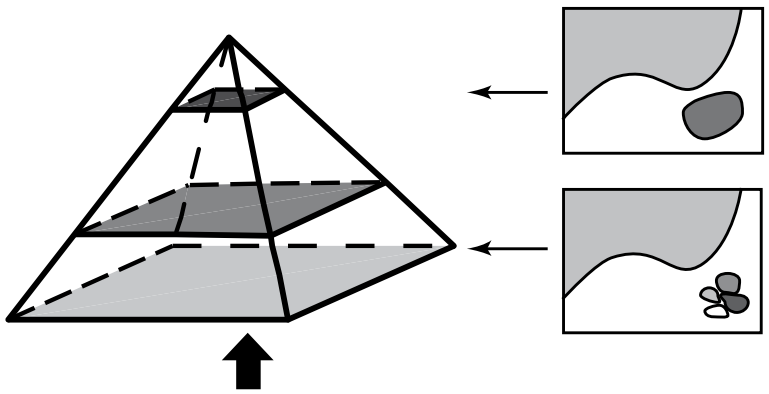


Figure 3-6 I-PWRC Hierarchical graph representation

The hierarchical segmentation operation starts from the bottom to register all fine details, while the higher level operation builds up on lower level outputs. In the related graph, the edges and vertices should be inherited from the lower level but weights of each edge need to be reconstructed due to the changes of the cluster size |**C**|. In this research, based on the complexity of the video contents to be analysed, the number of hierarchical levels can be varied from 5 to 25.

As the I-PWRC pushed up to the higher level, small regions are merged, which requires a dynamic mechanism for determining the k value denoted as $k(C)$. Since larger k values suppose to "trigger" the merging actions among larger regions, the $k(C)$ can be defined based on the largest region in the current hierarchical level, expressed as in Equation 3-21:

$$k(\mathbf{C}_i) = k(\mathbf{C}_{i-1})\left[1 + \frac{r \cdot \text{mean}(\mathbf{C}_{i-1})}{\max(\mathbf{C}_{i-1})}\right], \qquad \textbf{3-21}$$

$\mathbf{C}_i$ is the region collection on the i^{th} level in the hierarchical structure. The current k value is iterated and updated based on the region size of the lower level which is evaluated by the ratio, r, relating to the mean and the maximum region sizes. For keeping the most region details in the experiment, the initial $k(\mathbf{C}_0)$ and r were trialed and set at 0.17 and 2, respectively.

3.3. Feasibility Studies

As shown in Figure 3-7, a number of STV models were constructed based on self-recorded video clips to conduct the feasibility test on the innovative I-PWRC technique. These clips were recorded at the university campus and containing many

solid colour areas such as ground and clear sky, as well as many small areas with different textures such as tree leaves, window frames and cars.

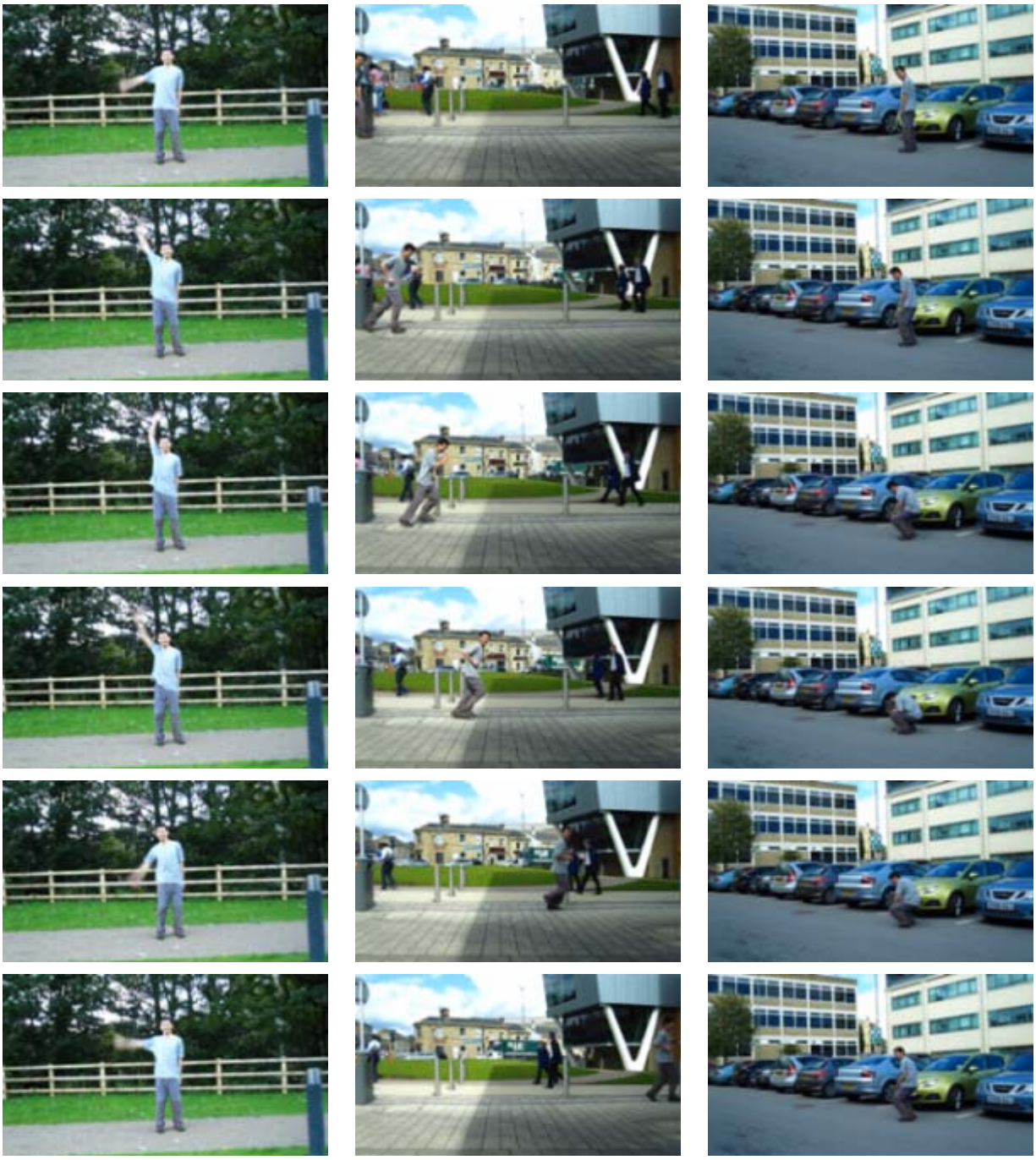


Figure 3-7 Snapshots of video clips for I-PWRC feasibility test

After the MS-driven pre-segmentation is filtered, voxels are grouped and clustered into small regions for graph initialisation. These small regions contain subsets of

accurate boundaries of the contents (3D shapes) in the STV model. In this experiment, the MS window size was set at 5 to 8 based on the complexity of the video inputs for extracting most of sub-region details. [Forsyth 2003].

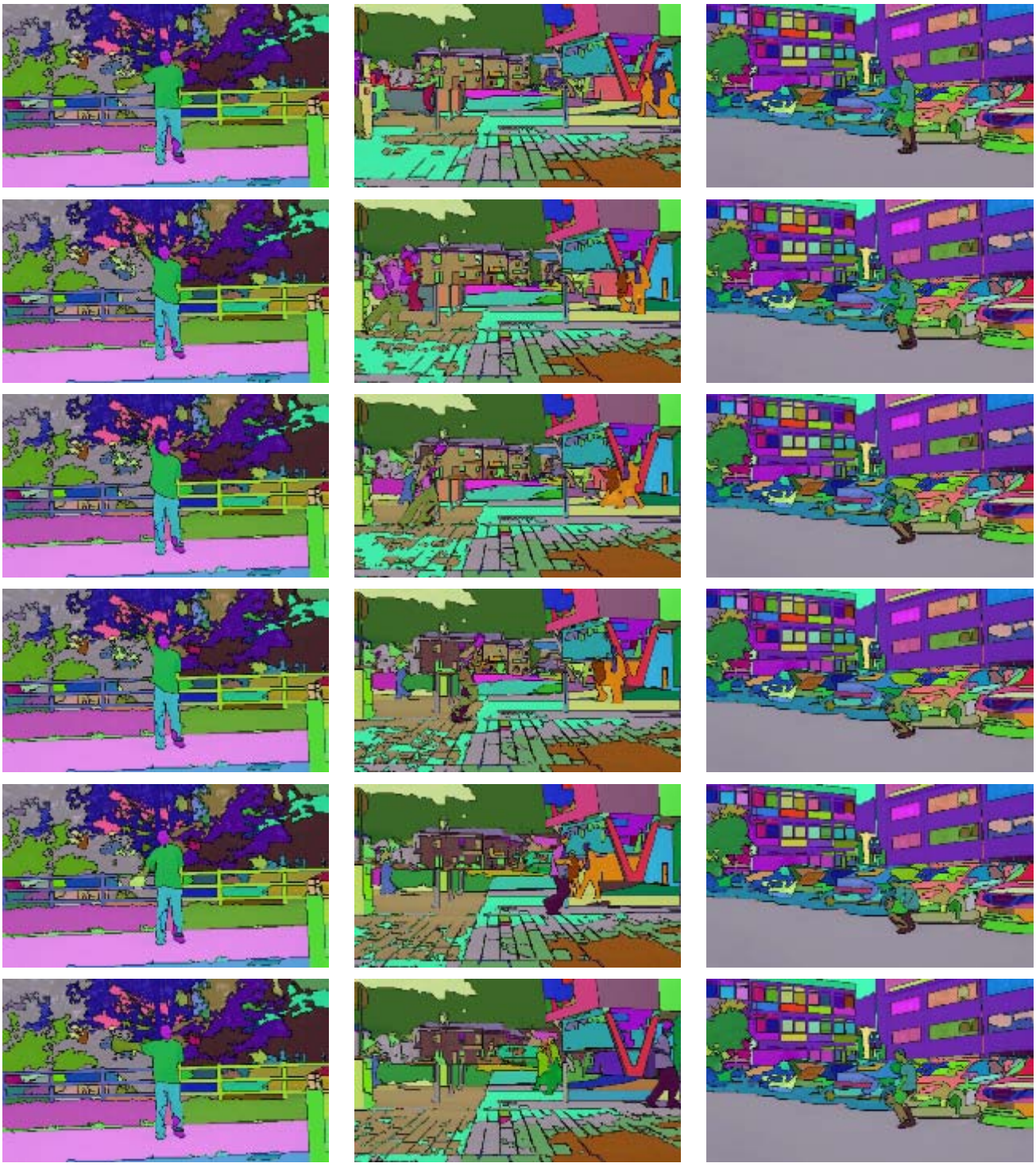
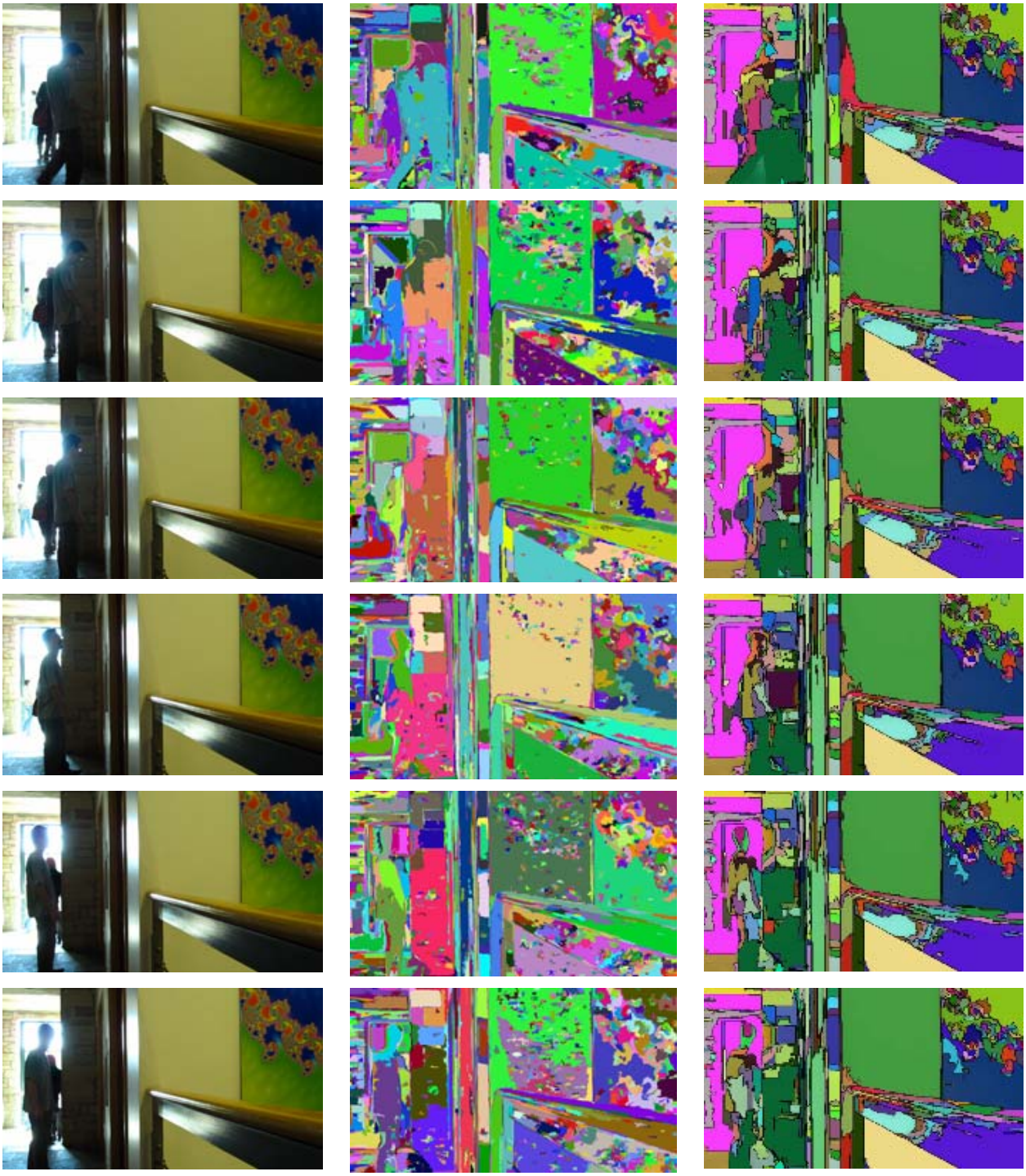


Figure 3-8 Sample segmentation outputs (see Figure 3-7) for original video inputs

The I-PWRC algorithm iterates in a hierarchical structure of 15-level during segmentation. The outputs are shown in the snapshots in Figure 3-8. It can be seen that most of the STV content boundaries have been found with even small textured areas highlighted accurately, which proves the design theorem of applying a dynamic $k(C)$ for adaptive handling of both large solid colour areas and detail textures.

Due to the 2D nature of the conventional PWRC, the 3D I-PWRC algorithms were also deployed on a 2D basis in this trial for comparative performance analysis. As shown in the Figure 3-9, the testing frames contain extremely bright and very dark regions, reflecting a substantial transition over a high dynamic range (HDR) illumination spectrum. The I-PWRC has shown its superior performance over the baseline algorithm on handling these scenarios. It is clearly visible in the Figure 3-9 (b), the segmentation outputs still contain a large quantity of small "fake" regions near large boundaries of walls and poster frames due to the rigid k value applied in conventional PWRC; while Figure 3-9 (c) has shown optimised results. It is worth noting that even in the areas subjected to slow illumination changes, such as the window and the wall in Figure 3-9 (a), the improved method still performs better in finding the correct boundaries since the dynamic k value renews the inner difference parameter *MInt* on each hierarchical level.

Close examination on I-PWRC segmentation performed in the trial still revealed regions where image/STV contents were over-segmented. The problem will be further tackled in the later stage of the event detection pipeline as covered in Chapter 4 and 5. More quantifiable tests and evaluations will be reported in Chapter 6.



(a) Testing Frames (b) Baseline 2D PWRC (c) Improved PWRC

Figure 3-9 Frame comparison between baseline PWRC and I-PWRC

3.4. Summary

In this chapter, high-level features inherited from raw STV datasets have been refined and clustered using an innovative I-PWRC volumetric segmentation technique. The outputs of this step which describing shape and texture video contents, such as boundaries and distribution of segmented regions, will be used as inputs for patterns analysis-based event detection in the video event processing pipeline. Based on the theoretical research and practical trial, the clustering-based segmentation strategy applied in this project has proven its effectiveness when extended from 2D to 3D. The I-PWRC algorithms developed in this research and their implementation utilised a number of key concepts and techniques including MS, PWRC, histogram descriptions and hierarchical pyramid data structures. The feasibility trial has recorded satisfactory and promising result. In the next chapter, the "fake" regions often caused by over-segmentation will be tackled in the event template matching stage.

Chapter 4. Volumetric Shape Extraction for Event Template Matching

Compared to traditional FBF-based video analysis strategies, a significant advantage from using the STV model is rooted in its distinctive ability to provide 3D geometric descriptions for dynamic video content features recorded in footage, which providing a theoretical foundation for event template matching. As illustrated in Section 2.5, video event features can be abstracted and recognised by using either global feature-based shape analysis methods; or local feature-based spatio-temporal interest point methods. A hybrid approach combined these features has been adopted in this research.

By integrating shape and feature point-based strategies, an innovative event detection framework has been developed that is based on the output from the I-PWRC operations (see Chapter 3) to feed to the shape-based event template matching (introduced in this chapter) before being refined by the spatio-temporal interest points-related operations (discussed in the Chapter 5). So the rest of this chapter is organised as following: Section 4.1 introduces a semi-automated template definition tool for event template composition. Section 4.2 presents the development of STV shape-based template matching algorithms through highlighting the efficiency and accuracy improvements from the volumetric operations. Analysis on the shape matching algorithm and its compatibility with I-PWRC is discussed in Section 4.3. Section 4.4 summaries the template matching work and addresses the issue of refinement.

4.1. Event Template Definition

STV-based template matching requires pre-built event templates being compared with event shapes segmented from video footages. An event shape might be a perfectly "cut" representing a video event of its entirety; or in a more complex way been "over-cut". Different strategies need to be formed to deal with these situations.

4.1.1. Template Matching Strategy Design

As illustrated in Figure 4-1, a "falling down" event is extracted and encapsulated in a volume structure that can be uniquely denoted by its geometric characteristics. These shape parameters can be used in geometric distribution analysis and shape matching in computerised and effective manner.

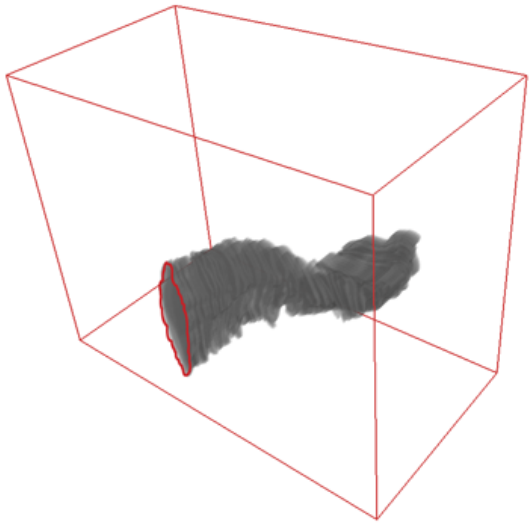


Figure 4-1 A "Falling down" event represented in STV model

As discussed in Section 2.5.2, many shape-based template matching algorithms, such as [Alper 2005], [Gorelick, Blank *et al.* 2007] and [Flitton, Breckon *et al.*2010], require "perfect boundaries" for functioning properly. But as explained in Section 3.3, meaningful video events are often difficult to identify and being separated from

uncontrolled backgrounds. Substantial numbers of fake regions may be generated along the way. For example, as illustrated in the Figure 3-8, the man in this video has been segmented into more than one part influenced by the texture of his clothes, which renders the "perfect boundaries"-based approaches invalid. In reality, these fake regions, named as over-segmented sub-regions, are common outputs from effort to separate signal from noise, especially in complex videoing environment.

To tackle this problem, various techniques, such as Region Merging have been proposed to improve the segmentation results. Although proven theoretically sound, computational overhead and latencies caused by these pre-processing steps had often prohibited the idea's practical usage [Wang *et al.* 2005]. In this project, a partial template matching technique has been devised and deployed for the STV-based event detection.

4.1.2. Forming Event Template

In a volumetric space, the matching process of a video event requires pre-defined 3D templates that provide geometries for comparison with 3D shapes retrieved from STV patterns. The template event shapes are representative models drawn from common or unique characters of a large group of similar events.

An ideal STV event template should be formed by accurate boundaries of event shapes that require artificial assistance during the template composing stage. In many 2D image processing applications, initial templates are generated by user-defined Regions of Interest (RoI). In the 3D volume space, STV event templates are defined by consecutive RoIs extracted in each frame in the event duration as shown in Figure 4-2. Because a typical single human action event lasts through 30 to 150 frames [Schuldt *et al*. 2004], the tedious RoI definition task needs to be alleviated in real

applications. In this research, a semi-automated event template composing tool has been developed for this purpose.

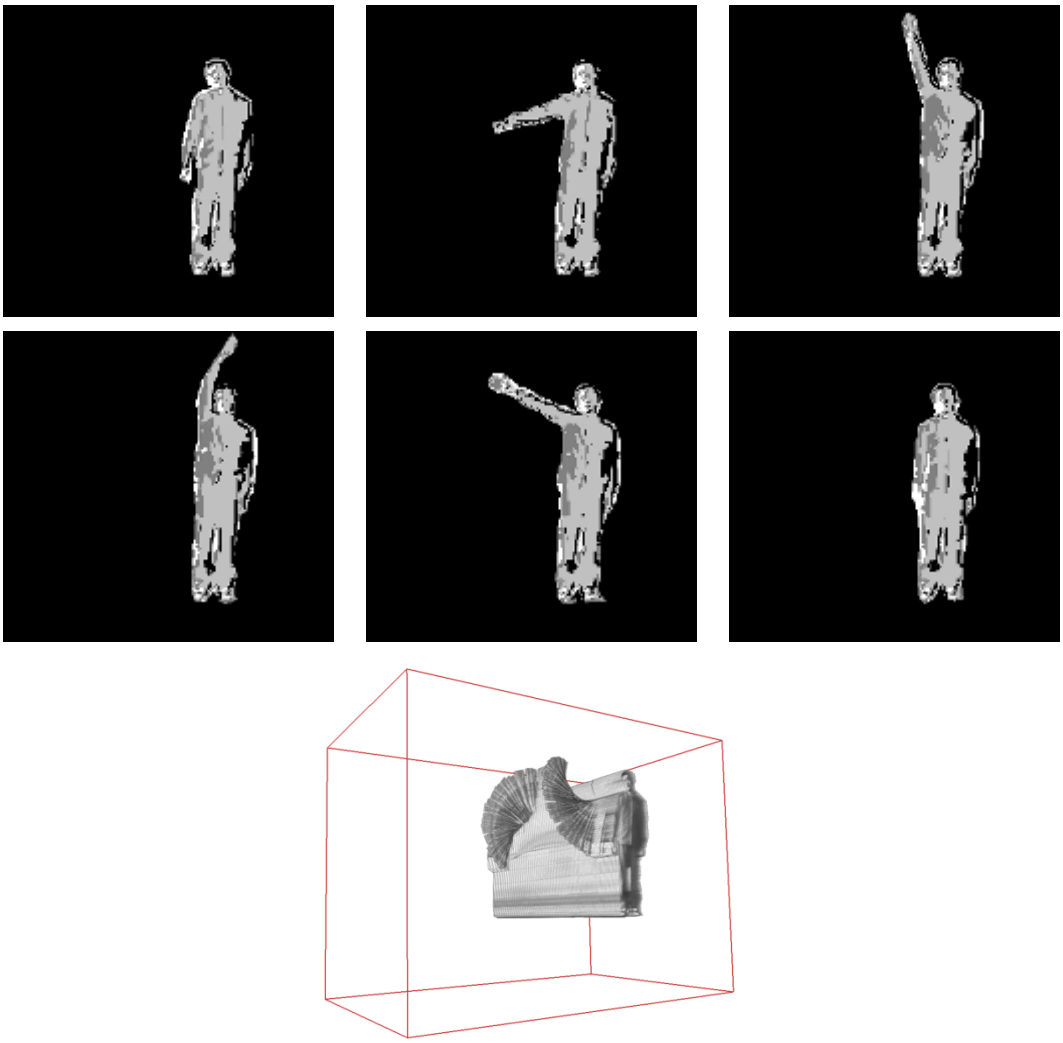


Figure 4-2 Building a "Waving" event template requires FBF-based RoI operations

Since the contours of event targets are drawn from consecutive frames, a straightforward method is to track the evolutional changes from previous RoIs in the immediate following frame and then renews the outlook of the contours using appropriate algorithms. An ideal set of algorithms for satisfying those acquirements are Active Contours (AC) techniques first introduced by Kass *et al.* [1988]. An AC algorithm combines model-based segmentation and tracking processes into a single operation.

4.1.3. AC Concepts

The AC algorithm implemented in this project is based on Kass' definition [1988] where an iteration process which ensures the so-called contour energy **E** becomes minimised. Pre-define a contour V with n points v_i in a row which is denoted as $V=\{v_1,v_2,\ldots,v_n\}$, where $v_i=(x_i,y_i)$ $i=1,2,\ldots,n$. The definition of the **E** on $v_i=(x_i,y_i)$ can be expressed as

$$\mathbf{E}_i = \alpha\mathbf{E}_{\text{int}}(v_i) + \beta\mathbf{E}_{\text{ext}}(v_i), \qquad \textbf{4-1}$$

The α and β are the weight constants. $\mathbf{E}_{\text{int}}$ is internal energy based on the shape of the contour and $\mathbf{E}_{\text{ext}}$ is external energy depending on the image Gradient around the point v_i. In addition, the internal energy reinforces the contour distribution and the grow/shrink tendency of a closed contour and the external energy drives contour points to the boundary of an object in the image. The output contour is decided by the minimum energy of these two factors.

As illustrated in Figure 4-3, **E** is an $m\times m$ matrix, which serves as a searching window in the operation. The energy of current v_i is located at the centre of the matrix and the rest of the matrix elements are energies of the neighbours of v_i. The location of minimum value in this matrix is the centre of **E** for next iteration, which means the location of v_i is modified to v_i' during the iteration by the current minimum energy. The iteration should be stopped only if $v_i = v_i'$.

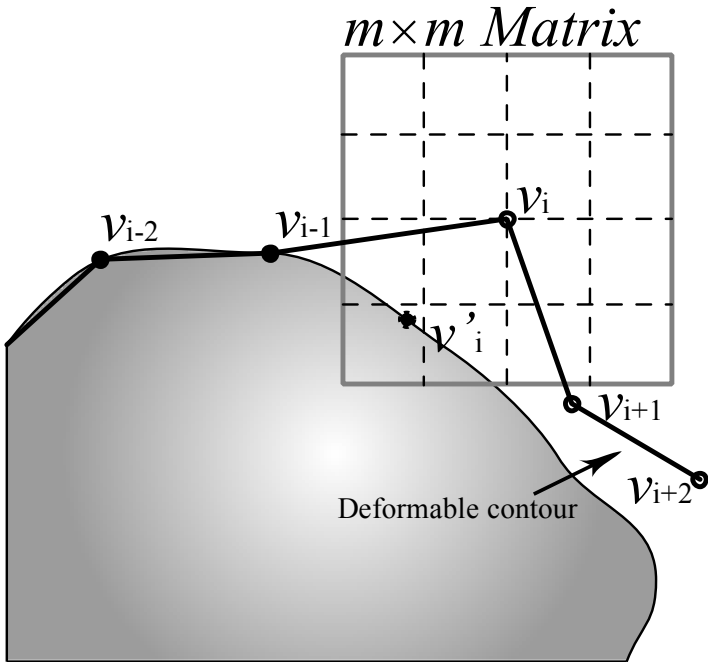
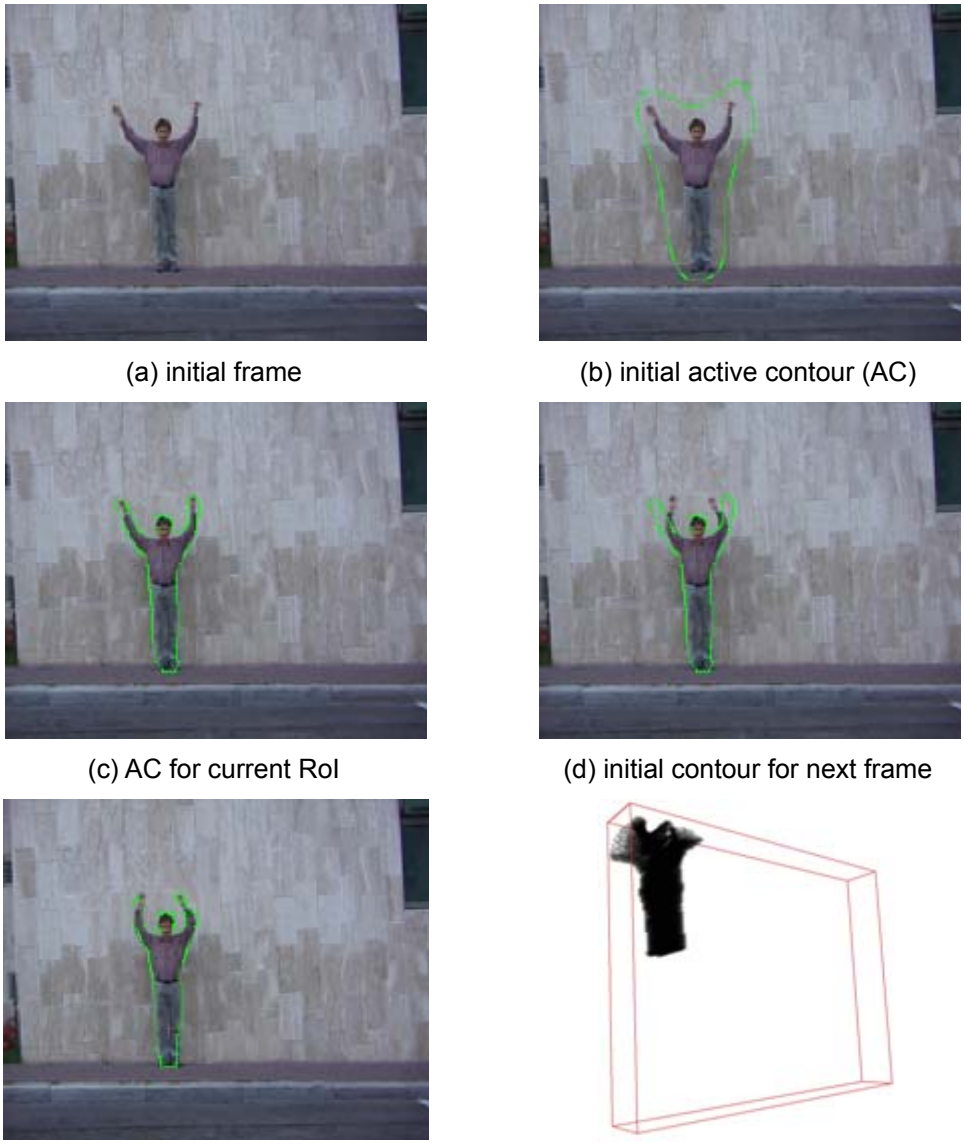


Figure 4-3 The illustration of the active contour algorithm

4.1.4. AC Implementation Principles

To ensure event detection proficiency, templates should be built with most representative event characteristics through carefully selecting video footages that should ideally only contain simple backgrounds and a single performer in each clip. Hence, the difference between background and event target is clearly distinguishable and easy to be identified by colour and intensity feature variables. In addition, STV event templates should be defined by "averaging" shape models created from multiple samples in each event category to improve the matching robustness under challenging real-world settings. Figure 4-4 illustrates the progression for defining a "waving" event template.



(a) initial frame (b) initial active contour (AC)

(c) AC for current RoI (d) initial contour for next frame

(e) AC for "next" RoI (f) composed STV event shape

Figure 4-4 Active Contour-based "Waving Template" Formation

After selecting a template video, the first frame of the video containing a specific human action event is manually initialised through marking the contour. Technically, the location and outline of the initial contour can be composed arbitrarily. However, the points defining the initial contour can only grow or shrink in the range of $m \times m$ matrix during each iteration. The time consumption can deteriorate rapidly if the initial contour is too far from the actual boundary of the event performer, which could also introduce risks of added noise and fake boundaries. For maintaining the

efficiency of the template definition operation, the initial closed contour needs to include the entire event target (the action performer) and the size should be approximate to the region of the performer, see Figure 4-4(b). After calculating and registering the first contour as shown in Figure 4-4(c), the contour output can be refined and maintained automatically or manually. The RoI of current frame serves as the input contour for the next frame (Figure 4-4(d)). At the end of this contour tracking process, the final RoI group is "stacked up" as an STV (Figure 4-4(f)) using the method introduced in Section 2.2.

Although the algorithm needs an initialised RoI at start and is often done manually, the accuracy of the output contour is generally sufficient to use for event template definition as proven in the experiment. In addition, the possible incorrect outputs, from each frame can still be maintained by operator intervention during the template definition process. The efficient semi-automated event template definition tool and editing mechanism had helped the construction of a set of high quality event templates in this project.

4.2. Event Shape Matching

4.2.1. Practical Issues

The I-PWRC segmentation outputs provide shape and boundary features for representing event profiles in the STV space. It transforms the event recognition operation into a 3D shape matching process. Conventional pattern matching techniques analyse the distribution of boundary segments directly based on the assumption that it contains "perfectly cut boundaries". However, as discussed earlier,

it is difficult and rare for video events to be cleanly separated from uncontrolled backgrounds. As a consequence, many "fake" regions can be falsely identified as interested regions. These small regions caused by over-segmentation are commonly treated as problematic and considered the main cause to the low efficiency of the relevant pattern analysis algorithms due to extra filtering required to "clean" the region boundaries.

In this research, the over-segmented volume boundaries are not viewed as "further improvement required" but an intermediate output that can be directly fed into the innovative 3D shape-based matching algorithm developed in this research. A close examination of the Figure 3-8 reveals that the over-segmentation has effectively identified all the intersected interested regions and identified all the minute shape boundary sections (sub-boundaries) of the volume. The proposed matching operation deploys a region filtering mechanism that directly operates on those segments through assessing a coefficient factor of the so-called Region Intersection Distance. Based on the early works from Ke *et al*. [2010], this approach can be classified into the "Region Intersection" (RI) matching category. One of the distinctive features of RI methods is their ability to perform shape-based event detection in challenging real-world setting where event signals are often immersed under complex background noises. The improved RI method devised in this research has explored the following design theorem.

4.2.2. Region Intersection Strategy

Rooted in Set Theory, RI-based shape matching calculates the differences (so-called "distance") between a pre-built volumetric event template and the segmented STV patterns. For example, if denoting A and B as two binary shapes, the distance metric

describing their similarities can be defined as $|(A \cup B) \setminus (A \cap B)|$. At run time, a pre-defined template set T will be "sliding" across the STV space for matching patterns. Considering an input STV model V containing an event located at $l=(x,y,t)$, the RI distance d can be represented as $d(T,V;l)$. Since V is an over-segmented model and all its sub-regions can be considered exclusive-non-overlapping. If V is composed of k sub-regions as $V = \bigcup_{i=1}^{k} V_i$. The overall distance between the event template and the detected pattern can be written as:

$$d(T,V;l) = \sum_{i=1}^{k} d(T,V_i;l), \qquad \text{4-2}$$

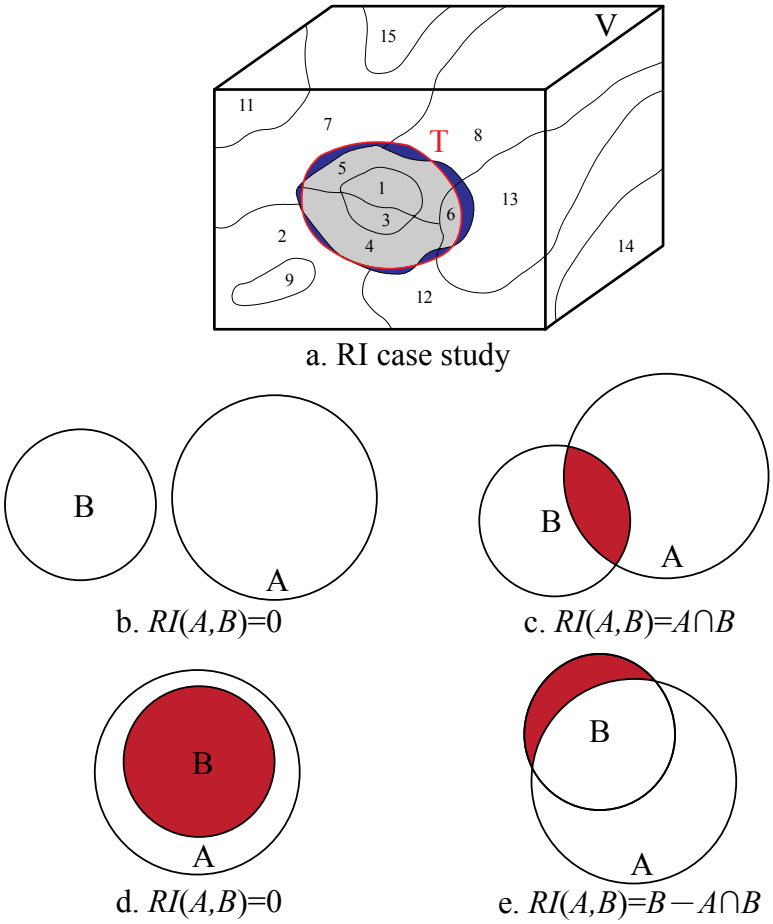


Figure 4-5 RI template matching algorithm and four possible scenarios

As shown in Figure 4-5 (a), the template T is highlighted in red and the over-segmented video volume V is composed of 15 sub-regions. The blue area highlight the

"distance" between V and T. Based on the contribution from each sub-region V_i, the blue area can be calculated according to the "intersection" rules defined as followings: If a sub-region is completely enclosed by the template or does not intersect with the template boundaries, such as V_1 and V_{11} in Figure 4-5 (a), then the distance can be defined as $d(T,V;l)=0$ (scenarios illustrated in Figure 4-5 (b) and (d)). Otherwise, the distance is defined as $d(T,V_i;l)=T\bigcap V_i$, an intersected area such as V_2 (scenarios defined in Figure 4-5 (c)). A more interesting case occurred as indicated by the sub-region V_4 in Figure 4-5 (a), where there is a large overlapped region with the template. In this case, the distance is defined as $d(T,V_4;l)=V_4-T\bigcap V_4$ - scenarios illustrated in Figure 4-5 (e). These cases in dealing with different types of sub-regions can be summarized as:

$$d(T,V_i;l)=\begin{cases}\left|T(l)\bigcap V_i\right| & if \quad \left|T(l)\bigcap V_i\right|<\left|V_i\right|/2\\ \left|V_i-T(l)\bigcap V_i\right| & otherwise\end{cases}, \qquad \textbf{4-3}$$

where $T(l)$ denotes template placed at location l.

However, deploying this strategy in the trials seems to cause high false positive rate especially when handing those small (over-segmented) regions, the RI distance can be normalised as:

$$d_N(T,V;l)=\frac{d(T,V;l)}{E_{\hat{T}}(\cdot,V)}, \qquad \textbf{4-4}$$

The normalisation factor $E_{\hat{T}}(\cdot,V)$ contains every possible template $\hat{T}$ that might be tested in the over-segmented volume V for RI matching, defined as

$$E_{\hat{T}}(\cdot,V)=\frac{1}{\left|\hat{T}\right|}\sum_{t\in\hat{T}}d(t,V), \qquad \textbf{4-5}$$

which can be simplified and approximated by $E(V)$ as

$$E(V) = \sum_{i=1}^{k} f(|V_i|), \qquad \textbf{4-6}$$

where

$$f(x) = \begin{cases} \frac{x}{2} - \frac{x}{2^{x+1}} C_x^{x/2}, & x = \text{even} \\ \frac{x}{2} - \frac{x}{2^x} C_{x-1}^{(x-1)/2}, & x = \text{odd} \end{cases}, \qquad \textbf{4-7}$$

Since the function $f(x)$ only depends on the size of sub-regions, the value of the normalisation factor can be therefore calculated independently and used as a Look Up Table (LUT) during the RI matching.

As discussed in Section 2.7, the variations of potential "event makers" can be substantial. By using the over-segmented STV sub-regions and RI matching, most image-based variations can be recorded and described comprehensively. After sliding the template set across the volume in a "searching window" manner, RI method is capable of marking all locations with a matching distance less than certain threshold, standing for a likely event.

4.3. Renovating the RI method

To address some of the implementation issues concerning the original RI method, this project has developed a number of performance augmenting techniques.

4.3.1. Improved Region Filtering

Original RI-based shape matching algorithms often compare a group of "connected" pattern segments with the entire template shape using the intersection rules. This

mechanism has shown its robustness when applied to "imperfect" segmentation outputs. Another function of the RI approach is its ability in combining over-segmented "small" regions into an "integrated" larger piece for a one-off intersectional test. This characteristic reduces the strict demands on initial video quality and format - a must-have for many other video processing approaches - as reviewed by Moeslund and Granum [2001].

However for 3D volume-based shape matching, the RI algorithm can be practically slow due to the amount of data to process and the redundant calculations involved on "non-contributions" regions. As indicated in Equation 4-2, at location *l*, the overall distance between an event template and a shape pattern is an exhaustive aggregation of the sub-distance from "1" to "*k*" standing for all the over-segmented sub-regions in a STV. It is clearly demonstrated by Equation 4-3, the sub-regions that are completely enclosed by or separated from the template boundaries will make no contribution to the effective distance. In practice, during a template matching operation, most sub-regions will not be intersected with a template especially when the over-segmented regions are much smaller than the event template. This fact is the root cause for the invalid looping operations and delays. To improve the efficiency of the RI-based process, a filtering mechanism has been introduced in this project to make early detection on invalid regions in a volume.

The improved RI approach is based on the over-segmentation outputs from the I-PWRC operations (see Section 3.2). In reality, any segmentation algorithms that preserve object boundaries are suitable to this practice, which takes advantage of the concept that any intersected regions must be "close" to the template location in a STV

model. Figure 4-6 illustrates a developed pre-processing system in this research to improve the efficiency of the RI-based template-matching operations.

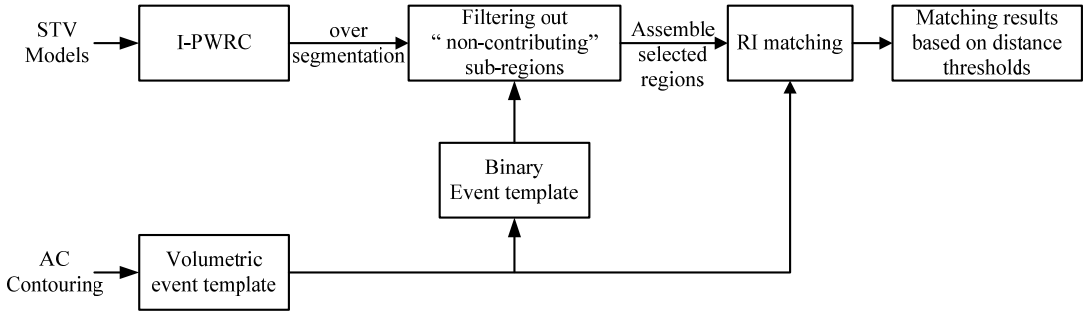


Figure 4-6 Region filtering pipeline

Figure 4-7 Binary form STV Shape of a waving template

The actual implementation of this step starts from defining a STV event template as a binary volume T_{bw} that contains "1s" for the silhouettes and "0s" for other voxels, as illustrated in the example in Figure 4-7. The boundary of the shape's surface T_{sur} can be calculated using 2D morphological boundaries extracted by

$$T_{sur} = \left| T_{bw} \oplus s - T_{bw} \right|, \qquad \textbf{4-8}$$

where s is the structuring element composed of 3×3×3 voxels with the $\oplus$ denotes the morphological dilation operation. In the volume, only the surface voxels are denoted

as "1s" for generating its surface counterpart T_{sur} that will be tested in each RI matching step at a random location l:

$$d(T,V;l)\big|T_{sur} = \sum_{i\in\widetilde{S}} d(T,V_i;l), \qquad \textbf{4-9}$$

where $\widetilde{S}$ is an "intersection list" accumulated from the filtering process over the entire sub-region sets (i=1,2,…,k) expressed as $\widetilde{S} = list(T_{sur},V) \subseteq S$. The *list* function, which is based on the geometrical distribution of T_{sur} in V, is defined as:

$$list\ \ (T_{sur},V) = filter\ \ (ascend\ \ (T_{sur}\times V)), \qquad \textbf{4-10}$$

where ascend(•) marks an ascending sort of the product T_{sur} and V at location l. The *filter*(•) stands for the accumulation of each of the non-zero entries from the sorted list. It is worth noting that once the relative positions of T and T_{sur} are specified in the candidate volume V, each V_i can be matched independently.

In this project, the filtering mechanism is automatically applied when a new position l is assigned at runtime. This simple process is efficient and only involves the morphological dilation, ascending sorting and some array operations.

4.3.2. Histogram-Verified Coefficient Factors

The RI operation introduced above can detect most event corresponding shapes in an over-segmented STV. However, the accuracy of this method can be further improved for real-world settings, through verifying the coefficient factor of RI distance before applying it for thresholding. As evaluated in Section 3.3, real-life video inputs usually contain both large uniform colour areas and small textured regions. The I-PWRC segmentation method can classify these contents in an over-segmented style consisted of both large and small sub-regions. However some extremely small regions around the large objects close to the real event shape boundaries can produce substantial

normalised RI distances, which leads to misidentification of the real event objects and a potential cause of the system's false-negative outputs. In the new approach, "rewards" have been designed to add "weights" to the larger sub-regions that effectively reduce the distance values. On the contrary, "punishments" have been given to the extremely small sub-regions intersected with the event template.

The evaluation scheme is automatically generated from a quantifying process on the intersected regions' local histograms to record the sizes and the numbers of the intersected sub-regions as shown in Figure 4-8.

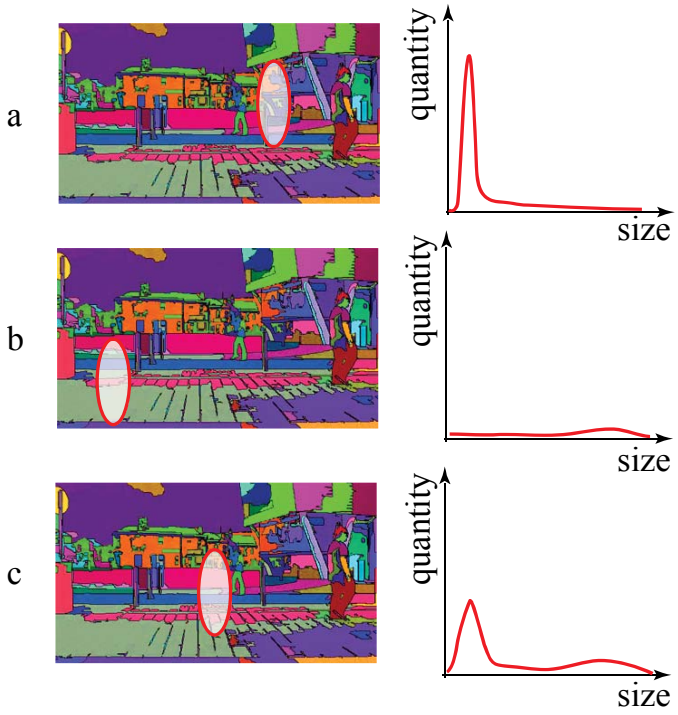


Figure 4-8 Local histograms used for evaluating the intersected regions

The local histograms discussed above can be used to compare with the benchmarking histograms inherited from the controlled and ideal RI matching scenarios. As illustrated in Figure 4-9, when the event actors and backgrounds are perfectly segmented, all feature points on the contour of the template will be closely coupled to

match to the segmented patterns, therefore the histogram will show a straight line lying on the horizontal "size" axis-effectively marking a zero distance.

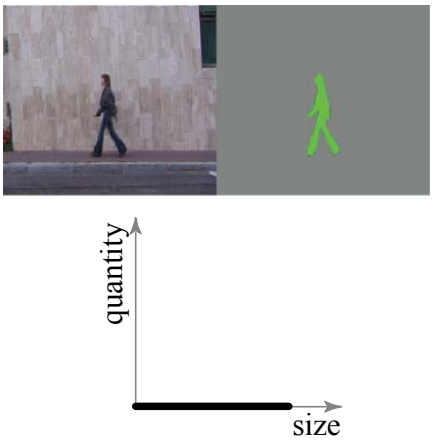


Figure 4-9 A Local histogram registering a "perfect" matching

For real world scenarios, there are three different situations while calculating the coefficient factors as illustrated in Figure 4-8 (a), (b) and (c), where the event templates are denoted by the artificial ellipses. In Figure 4-8 (a), the intersection parts are mainly composed by a large quantity of small sub-regions. The distribution histogram illustrated at the right hand side shows a single peak near the original point. In Figure 4-8 (b), the histogram is showing a largely flat curve with small fluctuations indicating fewer but larger intersected regional blocks. Figure 4-8(c) contains both large and small intersectional parts, where the smaller regions are in dominance; therefore the diagram shows a prominent peak in the histogram with smaller variations on other places. Through using the histograms, the distribution of different types of intersectional groups can be evaluated using the normalised χ^2 distance between the current histogram and a "perfect" one.

The coefficient factor can then be expressed as a linear transformation from the histogram distance, as denoted in Equation 4-11:

$$\tilde{d}_N(T,V;l) = d_N(T,V;l)\left[a + b \cdot C(T,V;l)\right], \quad \textbf{4-11}$$

where $C(T,V;l)$ is the normalised χ^2 distance of the histograms. The lower limit a and slope b control the degree of the correction of the RI distance. The value of the coefficient factor should be around 1, which is the threshold in switching between "rewarding" and "punishing". In the experiments, the range of changes is in between 0.6 and 1.4, which has been proven suitable for most of the video datasets tested.

4.4. Summary and Discussions

In this chapter, research work on event template definition and matching have been reported based on the I-PWRC segmentation outputs introduced in the previous chapter. One of the key techniques developed for this event detection stage is an improved Region Intersection-based shape matching method, which can handle the shape features generated from the over-segmentation operation. The baseline RI method adopted has been improved by a pre-filtering mechanism for optimising the input data. In the system design, the histogram distributions of the over-segmentation regions have also been used in the form of coefficient factors to verify the distance calculated for thresholding at run-time.

The research system of its current form can only handle video events that possess distinctive shape changes in the STV space. It cannot register motion occurred "inside" of a volume, for example, a front view of a human hand-clapping event. It is envisaged that the future works need to examine other modelling techniques, such as 3D articulated human body modelling for tackling the occlusion problems and to enhance the adaptability of the STV-based approach.

Chapter 5. Implementation Strategy and System Prototyping

STV shape-based event detection is a top-down process that involves event shape construction, segmentation and template matching operations. A modularised design of a prototype system has been adopted and been proved as a valid and effective approach. The event detection system prototyped in this programme has developed a process pipeline as illustrated in Figure 5-1. The functional modules (enclosed by solid-rectangles) denote the methodologies devised and utilised in the system. The modules marked by dashed-rectangles represent system optimisation techniques implemented. The system begins with a video signal acquisition module that generates STV models in the volume buffer. The I-PWRC segmentation then takes place on the models for extracting event shapes prior to the improved RI template matching operations. The benchmark event template formulation tasks are considered an off-line operation in this research.

In this chapter, the implementation strategies for improving system adaptability and performance robustness for challenging real-world scenarios have been discussed. Section 5.1 introduces a STV shape transformation algorithm and an adaptive template scaling scheme for handling model size and orientation problems. An efficiency-enhancement method for the RI matching procedures is discussed in detail in Section 5.2. Section 5.3 to 5.5 presents the system prototyping techniques and functional modules empowered by a unique data structure for handling the huge STV data files.

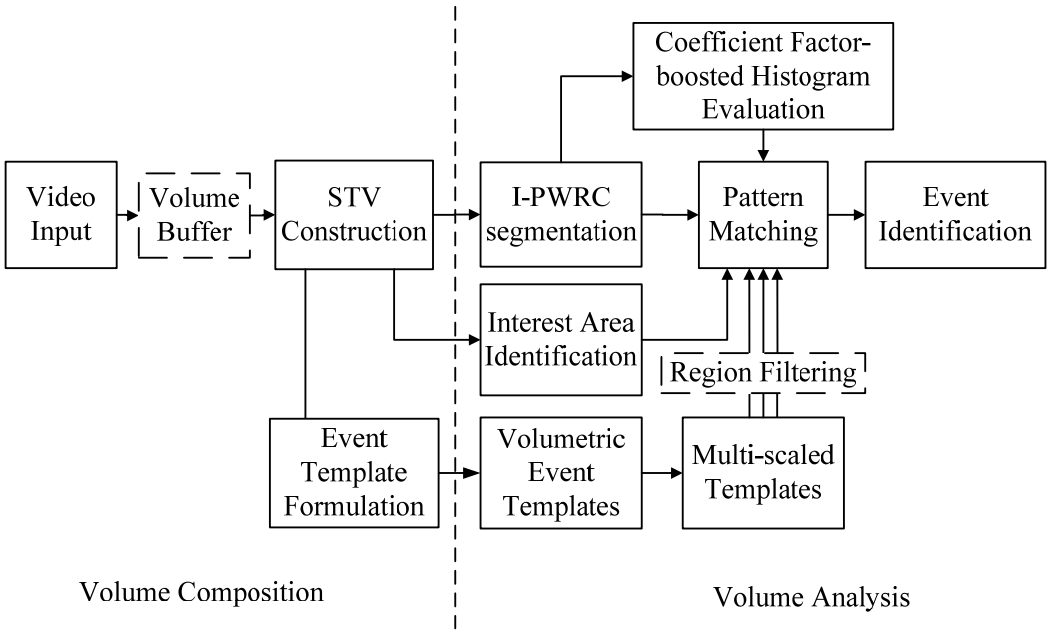


Figure 5-1 System pipeline

5.1. STV Normalisation

Video footage can contain events of arbitrary sizes, locations, and orientations in the scene depending on the camera positions. In addition, the same events can have variations in terms of their durations as they are subject to different event actors and video frame rates (see examples in Figure 5-2). Although the devised volumetric template matching approach possesses a certain degree of "tolerance" to those factors through threshold setting, the matching accuracy is still heavily influenced by system initialisation conditions. The system implementation strategy alleviates these uncertainties by introducing a "calibration" mechanism to normalise the segmented STV event models.

The “running” event template

Scaling variations

Orientational variations

Figure 5-2 Case for STV normalisation

Arguably, some local features can help to relieve the calibration/initialisation problems such as [Dollár 2005], [Niebles 2008] and [Bregonzio 2010]. However, those methods often involve time-consuming iterative or recursive computational machine learning processes. In addition, complex real-world settings often introduce strong noises. Therefore, local features-based techniques, such as BoW are difficult to deploy due to the time order problems and local features’ inherent sensitivity to noise signals [Wallach 2006].

In this research, the variations of the event model’s scales and orientations can be manipulated and standardised by the STV geometric transformation operations and achieving “normalised” models in the spatial and temporal domain.

5.1.1. Hierarchical and Multi-scaled Templates

One of the perceived solutions to solve the scale variation problem is to use the pyramid-like hierarchical templates that follow the linear model construction and re-scaling approach often used in image filtering and visualisation like mipmapping [Williams 1983].

The number of pyramid layers need to be selected carefully based on the adopted baseline RI matching techniques, which are often subject to the sensitivity of the RI

distance thresholds. Higher threshold values ensure lower false negative rates but limit the system's adaptability. In contrast, when applying lower thresholds, the shape matching outputs produce more positive results but also bringing in higher "risks" of high false-positive rate. For choosing the most suitable threshold, *Th*, for the RI template matching, the evaluation can be assessed by using the so-called "Receiver Operator Characteristic curves (ROC curves)", "Precision Recall Curve (PRC)" and Area under Curves (AUC) especially the AUC-PRC (see Section 6.2.1.).

Literally, the threshold range can encompass from 0 to $|T|$ (see Equation 4-3), the "relative threshold percentage" can be calculated by dividing the best threshold value by the $|T|$. Based on the theory introduced by Davis and Goadrich [2006], $Th/|T|$ is proportional to the system tolerance of the scale or orientation changes in a certain range, which also indicates the adaptability when introducing more re-scaled levels of the templates during the detection.

The tactic for deciding the number of templates needed in this system design is based on the following reference table, see Table 5-1.

relative threshold percentage （$Th/\lvert T\rvert$）	Number of re-sclaed templates	Linear re-scaled factors $(\vartheta_s, \vartheta_t)$
≥80%	5×5	0.6/0.8/1/1.2/1.4
60%~80%	3×3	0.75/1/1.25
≤60%	1×1	1

Table 5-1 Relations between the thresholds and the number of templates

Since the statistical distribution of the relative thresholds approximately obeys the Gaussian distribution, the entire range of the threshold has been divided into three sub-ranges based on the "three-sigma rule" [Fukelsheim 1994]. The upper/lower limitations of each re-scaled factor are based on the concept that "re-scaled template

should improve the detection accuracy" [Agarwal *et al.* 2004] compared with applying the original template based on the same $Th/|T|$. The number of templates is designed based on the Davis and Goadrich's conclusion introduced above and the fact of the system run-time performance during the experiments.

As indicated in the table, the number of templates in each threshold category contains a $n \times n$ grid of templates where n denotes the folds of scale changes in the spatial or temporal domain, where linear factors are denoted as ϑ_s (spatial) and ϑ_t (temporal). After determine the number of re-scaled templates, the linear re-scaled factors are deployed to the original STV templates along the spatio and temporal directions.

5.1.2. Normalising the Multi-scaled Templates

A problem for using the multi-scaled templates is that the RI distance varies when measured against different scaled templates and the matching outputs will be inconsistent to fixed threshold, for example, the RI distance can grow significantly if using a larger scaled template of an event.

This problem can be resolved by using a normalisation factor in conjunction with the coefficient factor (Equation 4-11) introduced in Section 4.3.2, expressed in Equation 5-1.

$$\tilde{d}_N(T,V;l) = d_N(T,V;l)\left[a + b \cdot C(T,V;l)\right]\frac{1}{\vartheta_s^2 \vartheta_t}, \qquad \textbf{5-1}$$

where the linear re-scaling factor ϑ can be referred as normalisation parameters in the spatio-temporal domain.

5.2. RI Interest Area Identification

Through sliding a searching "window" across a digital image to identify per-defined patterns is a widely used technique for automatic recognition, such as [Papageorgiou 1998] and [Viola and Jones 2004]. But it was also made clear in this research that the time consumption of using sliding "window" for huge data involved in the STV structure is grown magnificently.

Recalling the previous discussions on single human-based action event definitions (see Section 2.5), it is safe to presume that those events are mainly caused by dynamic objects. After removing the static areas from a STV by filtering, the searching range can be narrowed down on these "interest" areas for event detection, rather than exploring the entire 3D cube. In this research, the interest areas are defined by a group of 3D feature points, whose locations and distributions along the temporal axis in volume space indicate the possible event-occurring areas

In STV space, the dynamic information can be represented by the non-linear trajectories of the interest points with the length of a trajectory denoting the "duration" of an event (of events).

5.2.1. Locating the Interest Feature Points

In the classic computer vision research domain, the interest feature point theory is often applied when describing image contents using a number of 2D points to "mark out" the distinctive regions within an image. These representative regions often stand for edges, corners, or other spatial features containing certain distinctions in the image. Frequency domain filtering can often facilitate the enhancement of those features. Since the features are based on pixels and their neighbours, the interest feature point method is considered a local feature-based technique.

The 3D interest point extraction method developed in this programme is simulated by Lowe's Scale Invariant Feature Transform (SIFT) [2004], which can abstract scale, rotation and illumination independent image features from 2D patterns. The SIFT feature abstraction algorithm is illustrated in Figure 5-3. After two convolutions on each scale level by using the Gaussian kernel [Lindeberg 1994], "Difference-of-Gaussian"(DoG) image pyramids can be composed for extracting feature points. As illustrated in Figure 5-4, the SIFT feature is determined by comparing the candidate pixel (marked by "dot" in the figure) with its neighbours on the current, the higher and the lower scale levels. Based on the "cascading" principles [Viola and Jones 2004] and the key point localisation filters [Lowe 2004], the "maximum" or "minimum" pixels are denoted as SIFT feature points in the pyramid.

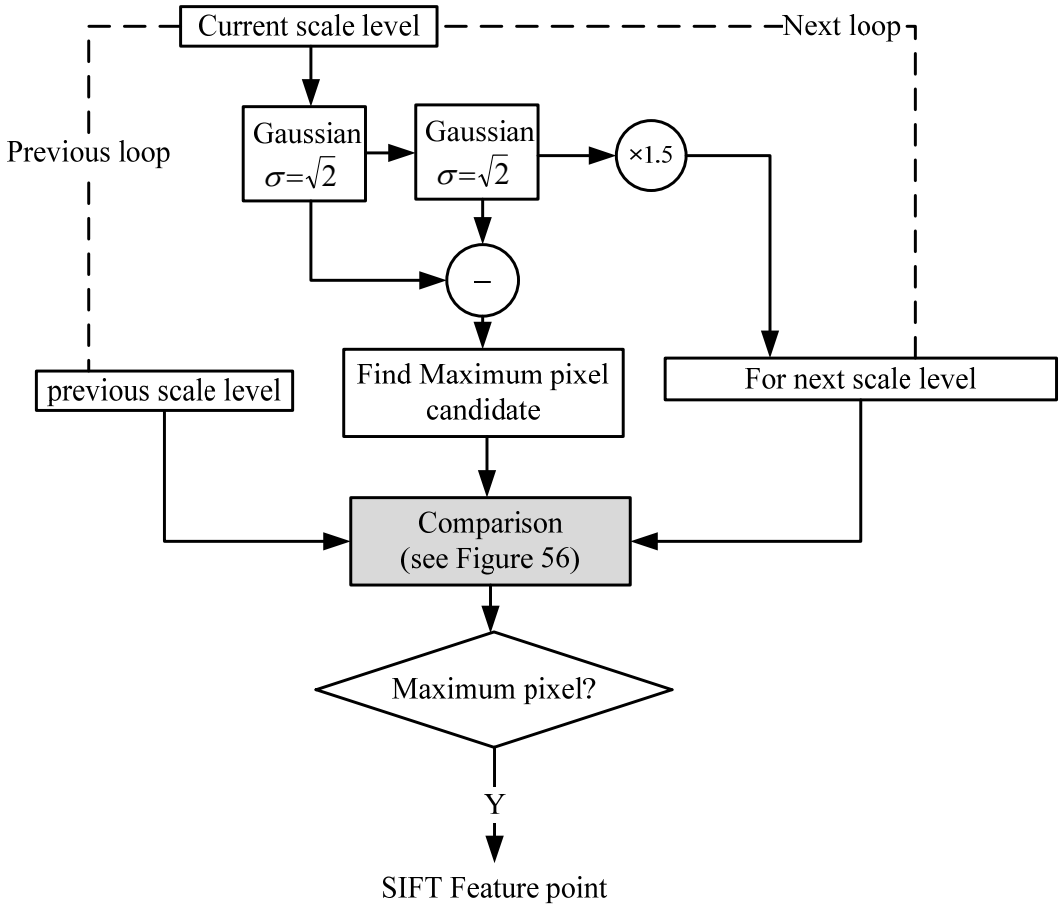


Figure 5-3 SIFT process flowchart

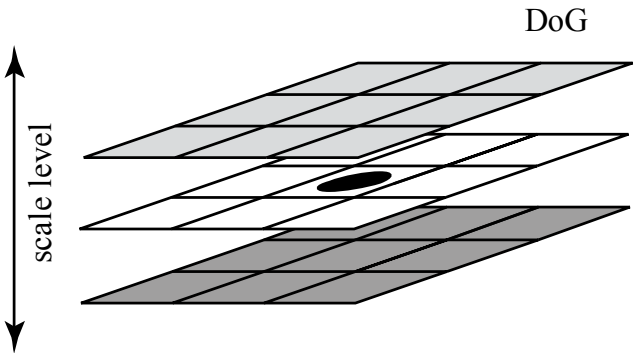
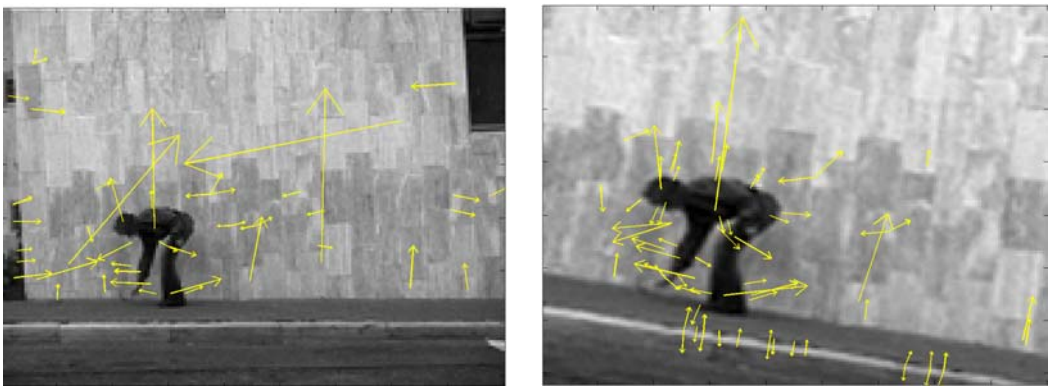


Figure 5-4 SIFT candidate comparison

Figure 5-5 (a) shows an SIFT feature extraction example. In contrast, Figure 5-5 (b) shows the same snapshot in (a) subjecting to a rotation of 10 degrees and reduced brightness of 20%, as well as the horizontal factor of 4.5 and the vertical factor of 2.9. An empirical study seems to show that most of the feature points marking the human body have been preserved during the transformation indicating a robust feature point extraction performance.



(a) (b)

Figure 5-5 SIFT Features

The SIFT method can be directly applied to STV data. A number of research groups have piloted the STV-based 3D SIFT feature extraction algorithms, such as [Dalal and Trigges 2005], [Lopes 2009] and [Flitton 2010]. In this research, SIFT features are extracted from the 2D XY, XT and YT planes along each of a STV model in a cyclic

fashion to reduce the time consumption of the algorithm's 3D counterpart. Similar as SIFT features applied on 2D image plane, the interest points abstracted from XT and YT planes of STV contain geometric transformations - and illumination - independt features which also sensitive to the "high frequency" area such as corners and edges. Due to the temporal information are naturally involved in these plains, the "corners" and "edges" are actually produced by moving objects of videos. The possible event area can be located based on the coordinates of SIFT features, which is sufficient to satisfy the needs in the proposed system pipeline for locating the RI interest areas.

5.2.2. SIFT-based Interest Area Formulation

The SIFT feature points extracted in the previous section are used for tracking and evaluating the 3D trajectories in the STV space. The tracking approach follows the principles of the typical frame-based flow tracking algorithm introduced by [Horn and Schunck 1981]. The "stable" trajectories with the life span surpassing 300ms in this design will be kept for further processing.

Equation 5-2 defines the Interest Area (IA) in this research based on the SIFT feature point set **p** and their related trajectories $\mathbf{T_P}$:

$$S = \bigcup_{i=1}^{n} s_i(\mathbf{l}, \mathbf{r}), \qquad \textbf{5-2}$$

where S denotes the distribution of the interest areas for the searching window composed of n element regions (S_i)); n is equal to the quantity of SIFT points belonging to $\mathbf{p}_i$. **l** and **r** represent the location and range consecutively that

$$\mathbf{l}(x, y, t) = \mathbf{p}(x, y, t), \qquad \textbf{5-3}$$

and

$$\mathbf{r} = \text{square}(\mathbf{T_p}), \qquad \textbf{5-4}$$

In the Equation 5-4, the range of interest areas are defined by a series of 2D squares whose barycentre is located by Equation 5-3 in the XY-plan of the STV space. The 2D squares can form a 3D "tube" along the STV's temporal axis. As illustrated in Figure 5-6, the tendency of the tube is largely conforming its enclosing SIFT trajectories. Based on the experiments carried out in this research, for keeping the most effective and accurate system run-time performance, the size of square, denoted as *s*, has been trialed and classified into 3 pixel-level categories based on the average event time durations: 5×5, 9×9 and 15×15, that are associated with the trajectories length *l* falling into the ranges of 300ms≤l<1000ms, 1000ms≤l<2000ms, l≥2000ms, respectively.

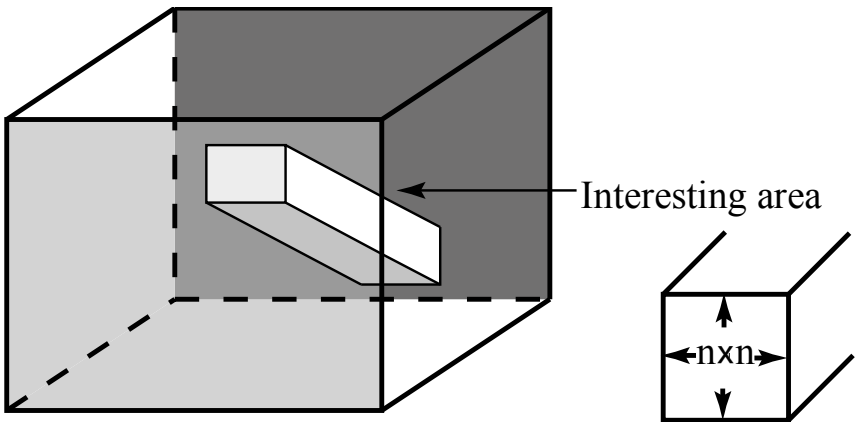


Figure 5-6 3D Interest area construction

In the feasibility experiments, it was observed that the time consumed by this extra feature processing step was approximately 120ms – a small latency in return for significantly improved overall system efficiency. Through identifying and locating interest areas from the entire STV models, RI searching window can be readily located on the high probability regions before applying the RI matching steps, see Figure 5-7.

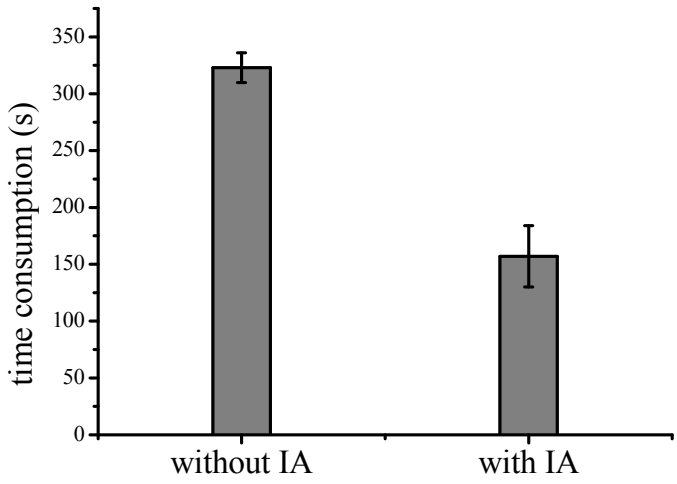


Figure 5-7 Efficiency improvement from interest area identification

5.3. System Modularisation and Data Pre-processing

The investigation on STV-based video event detection and system prototyping in this research focuses on three aspects: rapid and lean STV model construction, adaptive feature segmentation, and optimised shape-based template matching.

(a) (b) (c)

Figure 5-8 System hardware platform

As illustrated in Figure 5-8 (a), the system prototype has been setup on a host PC equipped of an AMD 2.62 GHz Athlon CPU with 2G RAM. The video signals are captured by a consumer grade CCTV camera (Figure 5-8 (b)) and a PCI analog image data acquisition card (Figure 5-8 (c)). The device algorithms in this research were initially tested on the simulation tools such as LabVIEW and Simulink with extensive use of MATLAB and OpenCV 2.2 programming APIs and functions.

5.3.1. Data Filtering Consideration

The challenges of this research at the data acquisition stage can be concludes as two aspects. Firstly, original STV models are often extremely large in data size due to their volumetric 3D nature and rich per-voxel characteristics. Secondly, the volume data processing techniques are generally inherited from 2D segmentation algorithm such as the PWRC method introduced in Chapter 3, which involves many looping and branching programming structures that can introduce serious latency to the computational efficiency. These problems often hampered the effort in the past in adopting the STV-based methods in real-world settings, especially for those time-critical applications.

A straightforward approach for relieving the negative impact from the huge STV data size is to reduce the total voxel quantities that need to be handled by relevant algorithms. This data reduction operation can be performed by specific software engineering techniques such as data reconstruction, compression, and pre-filtering to remove the so-called “non-contribution” or redundant voxels at the initial STV construction stage. In this research, the optimisation technique developed and deployed for this purpose was based on an “on the fly” computer memory management strategy called volume buffering.

5.3.2. STV Buffering Technique

Live video inputs vary greatly from sensors and video codec (that is data format and compression style adopted) parameters such as resolution, colour setting and video length. It is both unnecessary and impractical to transform an entire video file into a single STV volume. A solution proposed in this research has adopted an "on the fly" or procedural mode to generate fixed-length transitional STVs based on the pre-defined event durations before pushing them into the queue-like process pipeline.

At runtime the system starts with building a buffer (assigning memory) to the incoming video stream. The index of the first frame and the size of a STV model are customisable and dependant on the pre-defined action template sizes. Once the last event matching step is completed, the buffer assigned for holding the STV model will be freed from the memory to avoid accumulating memory footprints for the next cycle.

The benefit of this design is achieved through harnessing an efficient computer data structure - queue - that enables a first-come-first-serve operational order and its intrinsic flexibility in handling arbitrary sizes of data packets. Figure 5-9 illustrates the STV construction and registration operations engaged in this design. Currently, the assignment of the starting frame's index number has been simplified by halving the previous STV chunk (except the first "on-fly" STV which starts from frame No.0) and using its "middle" frame as the beginning for the next model. A more robust sampling approach, for example a one-third start from the previous STV or even an arbitrary starter, should see a more adaptable reconstruction process with an improved chance to encapsulate an event occurred but this will have longer latency. As shown in Table 5-2, this process has been implemented as shown in the pseudo code, where

the setting of the STV size and the starting position has been parameterised for improving the process's adaptability.

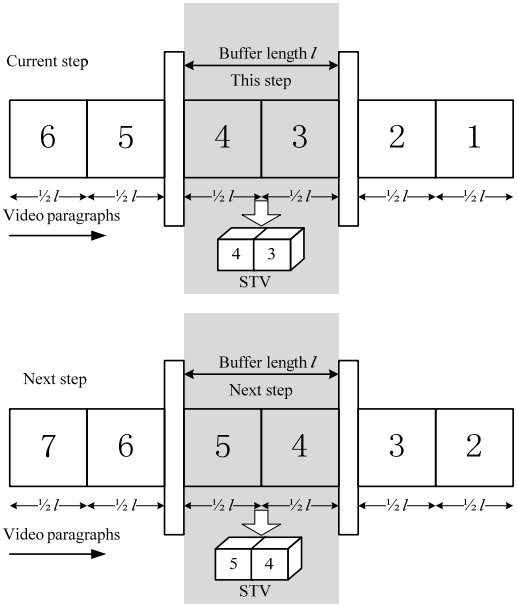


Figure 5-9 Volume buffer procedures

Pseudocode volumeBuffering (input videoFile)

START:
//Initialization
Allocate appropriate buffer size "***L***" based on videoFile configuration;
Calculate number of time-tablets "***I***" based on event template durations;
Calculate the remainder frames "***N***" at the end of the videoFile;

//Traverse through entire the video volume
Loop (***I***)
    //Calculate the new starting point and length "***R***" of the input STV
        if (at the end of video)
            R = ***N***;
        else
          R = ***L***;

//STV-based template matching
    Release tested STV;
    Compose and renew STV;
    RI matching;

END

Table 5-2 Volume buffer Pseudo code

5.4. STV Feature Extraction

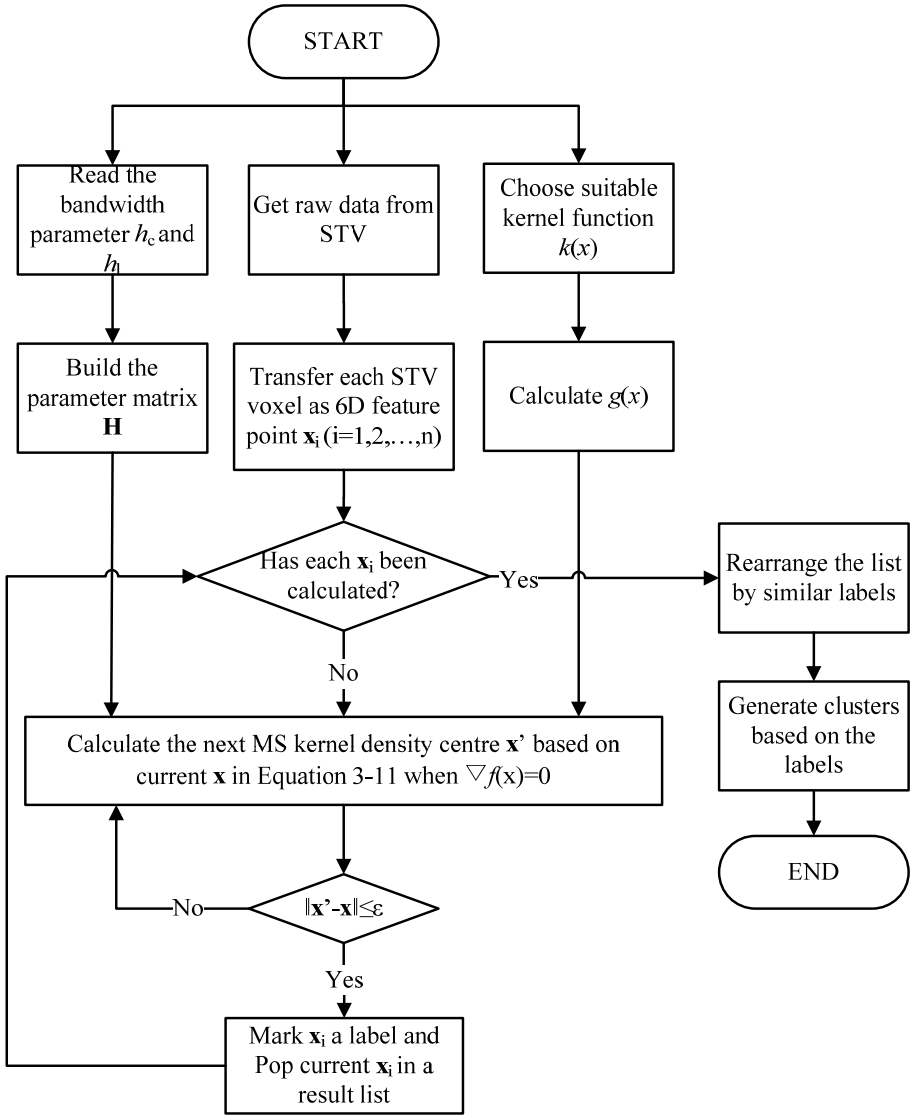


Figure 5-10 The flowchart of the MS pre-segmentation algorithm

- MS-driven Pre-clustering

When applying MS in 2D segmentation and clustering, the inputs from an image model are often restricted to the space coordinates and the colour values of all the 2D pixels. In turn, the defined feature space is of 5 degrees of freedom (x,y,r,g,b), in which (x,y) denotes the space coordinates and (r,g,b) represents the colour of a pixel.

In the case of a 3D STV model, its feature space naturally extends to 6D as (x,y,z,r,g,b), where (x,y,z) denotes the space coordinates of a voxel and (r,g,b) for the relevant colours. Therefore, a multivariate kernel density estimation process similar to its 2D counterpart can be implemented in the order as shown in Figure 5-10 (refer to Equation 3-11 in Section 3.2.1)

In this STV pre-clustering process, an iterative operation is carried out on each feature point that shifts the current feature density centre to the densest local region controlled by the bandwidth parameters. This design effectively reduced the graph vertices for the following I-PWRC processes.

- Composing histogram-based region graph

The region description words for I-PWRC segmentation have been implemented as four distinctive functions in the system as illustrated in the Figure 5-11, these functions handle the tasks of colour space transformation, region histogram representation, histogram distance calculation, and graph composition, respectively.

Based on the Equation 3-13 to 3-17 in Section 3.2.2, the colour space of each voxel can be transformed from RGB to L*a*b*colour domain. This colour information is in turn used for defining STV region textures inherited from the MS pre-clustering results in the form of local histograms, which are literally the vertices of the constructed region graph. After calculating the histogram distances based on Equations 3-18 and 3-19, the graph can be saved as the so-called Dictionary of Keys (DoK) for the following I-PWRC operations. It is worth noting that the colour histogram of many small regions can consume as much system memory as larger regions, especially at the early iterative stage of I-PWRC. To alleviate this problem, in

the design, the histograms are modelled as a 2-column-table containing all the non-zero L*a*b* vectors.

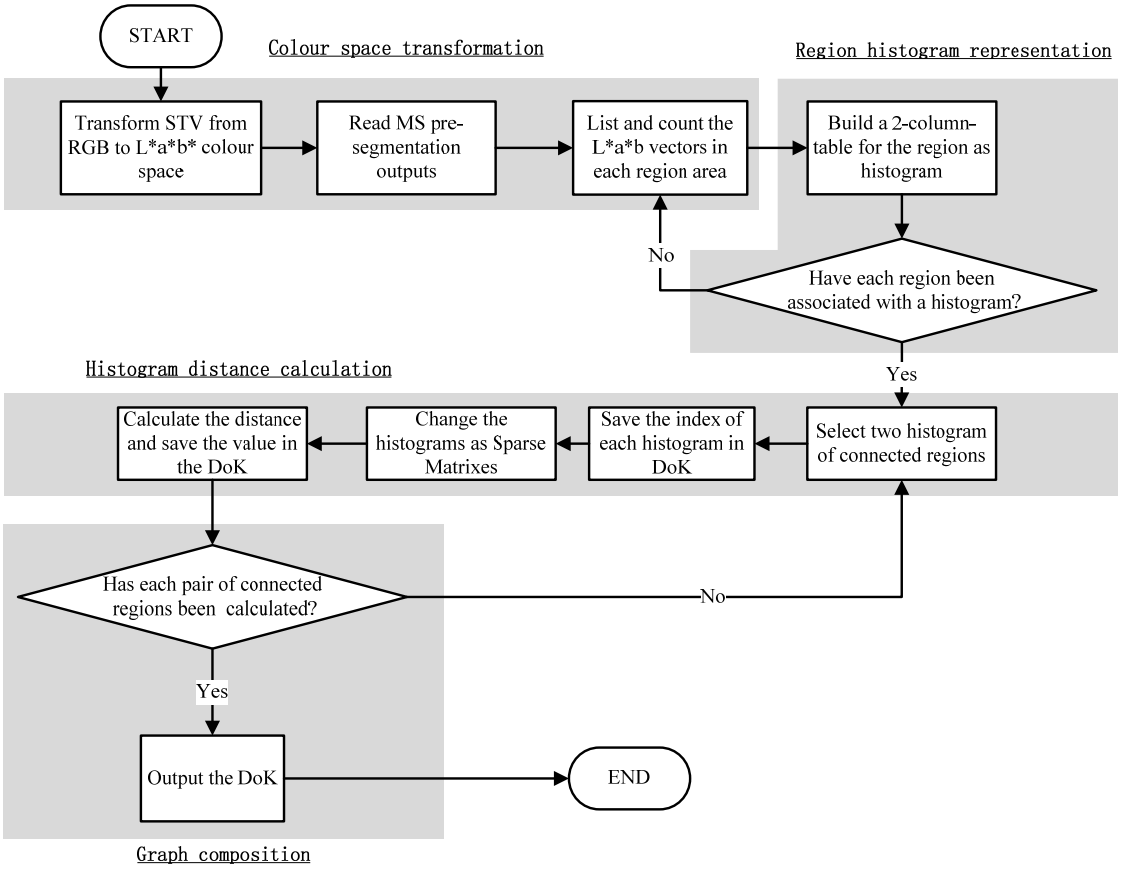


Figure 5-11 Constructing histogram-based region graph for I-PWRC

- I-PWRC Development

Based on the above discussion, the I-PWRC method devised in the research has been implemented in a more reliable and efficient manner for dealing with STV models. Table 5-3 provides the pseudo code for the I-PWRC operations engaged in the process pipeline.

Pseudocode Improved Pair-wise Region Comparison

INPUT
1. Spatio-temporal Volume
2. Mean Shift window size factor: h_c and h_l
3. Initialised Pair-wise Region Comparison factor $k(\mathbf{C}_0)$
4. Hierarchical levels n

OUTPUT
STV with Labelled segmentation regions $\mathbf{C}_{n-1}$

ALGORITHM
Initialise $\mathbf{C}_0$ with STV-based Mean Shift segmentation. (Equation 3-11 and 3-12)
Transform STV colour space from RGB to L*a*b*.(Equation 3-13 to 3-17)
Loop i from 1 to $n-1$
    Build histogram for each region in $\mathbf{C}_{i-1}$
    Represent $\mathbf{C}_{i-1}$ as graph:
        The vertexes value is L*a*b* colour
        The weight of edges are Cha's minimum histogram distances. (Equation 3-20)
    Calculate $k(\mathbf{C}_i)$ (Equation 3-21)
    Calculate $\mathbf{C}_i$ based on original PWRC method (Equation 3-3)
    Build next hierarchical level on lower resolution (Except the last loop)
End Loop
Output $\mathbf{C}_i$

END Pseudo code

Table 5-3 Pseudo code for the I-PWRC method

5.5. Event Shape Matching

- Region Filtering Implementation

As discussed in Section 4.3.1, the purpose of the improved region filtering algorithm is to generate a candidate list in the form of a 1D array containing the labels of the intersected regions from the I-PWRC segmentation outputs. Figure 5-12 illustrates the algorithm and its key operational functions that are further clarified in Figure 5-13 through highlighting the core processes empowered by Equation 4-8 to 4-10 explained in Section 4.3.1.

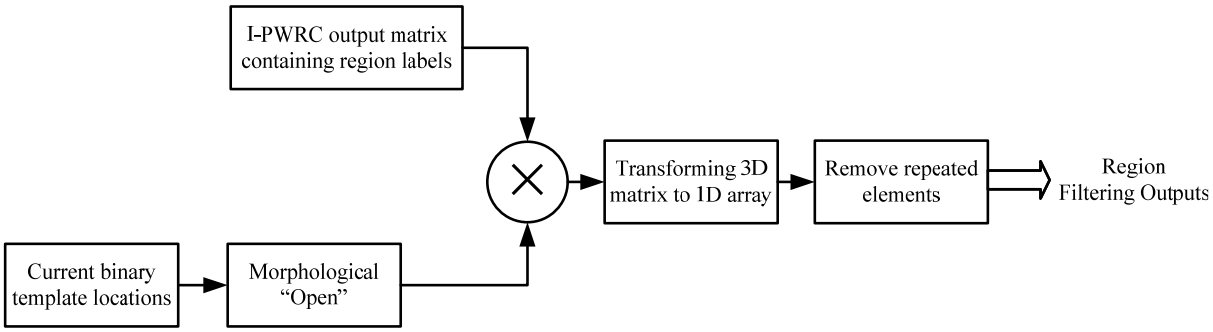


Figure 5-12 Region Filtering algorithm

V	1	1	5	5	5	7	7	7	3	2	2	2	4	4	6	6
T_{bw}	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0
T_{sur}	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0
$T_{sur} \times V$	0	1	0	0	0	0	0	0	0	2	0	2	0	0	6	0
$ascend(T_{sur} \times V)$	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	6
$filter(\bullet)$	1	2	6													

Figure 5-13 Processes for filtering out the "non-contributing" sub-regions

Recalling Figure 4-5, the I-PWRC produces many regions that are marked with the "non-contributing" label (the "0"s) in the array. Another optimising strategy applied in the programme is to form a "linked list" rather than directly using the 1D array for representing 3D matrices, which will omit the "0"s during the list initialisation, hence reducing the overall data size.

- Histogram-verified Coefficient Calculation

The improved RI algorithm has been implemented in the style indicated in Figure 5-14, excepting the main morphological operations involved, the Coefficient Factor-driven verification operations for the RI distance as shown in the right side of the flowchart

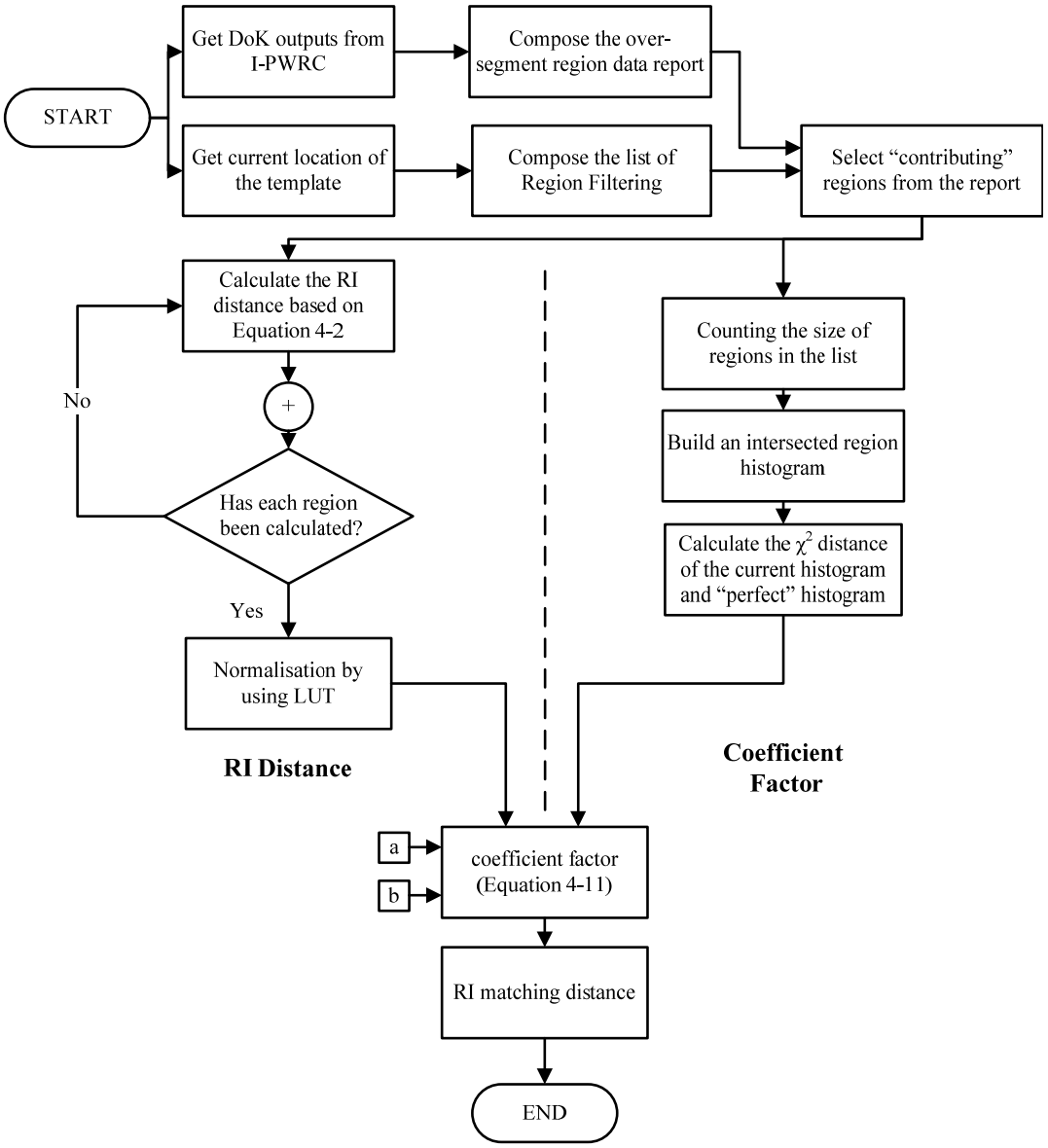


Figure 5-14 STV-based RI matching algorithm

It is worth noting that the DoK imported from I-PWRC can also form a Linked-List of a structures (STRUCT in programming term) containing the region label indices and the distribution of the over-segmented regions. This structure has been used as a "data report" in the system prototype for calculating the RI distances and the histogram-verified coefficient factors during template matching. As indicated by the experiment results detailed in the following chapter, this programming approach is more effective in maintaining the STV features.

5.6. Summary and Discussion

In this chapter, a number of practical optimisation techniques have been introduced for improving the robustness and efficiency of the proposed event detection system. The multi-scaled template mechanism has been used to improve the shape matching performance based on the RI distance transformation. After combing the normalised parameters with a coefficient factor, the RI distance can be adapted to spatial and temporal changes of the event performers in real applications. The overall system efficiency has also been improved by using the optimisation techniques such as "interest area" and "volume buffering" for controlling data sizes. The former method is a filtering operation that "predicts" the areas with high probability for containing the targeted events. The latter technique relies on an innovative buffering data structure for improving the runtime performances, hence enhancing the system prototype's suitability for handling arbitrary video file sizes based on better computer memory management.

Chapter 6. Experiment and Evaluation

The runtime performance of the devised event detection techniques and a system prototype have been tested and analysed in this chapter. Following the order of the process pipeline, the precision and efficiency of the STV-based shape matching processes have been assessed and benchmarked against other classic systems and algorithms. Various popular colour spaces and RI template formats have been examined to highlight the proficiency of the newly developed feature extraction and pattern recognition techniques. The evaluation on the overall system performance has been focusing on the robustness of its modularised design as detailed in Chapter 5. Both controlled and uncontrolled video settings and inputs have been applied in the experiments to assess the system's adaptability and robustness for real applications.

6.1. Test Data Acquisition

Several public datasets have been utilised in the experiment designs for their popularity in CV research areas and benchmarking potentials. Among those tested datasets, 2 open-access online video library sources and a self-made one have finally been adopted for relevant experiments due to their representativeness underpinning distinctive characteristics as listed in the Table 6-1. Selected frames taken from these datasets are illustrated from Figure 6-1 to Figure 6-3, respectively.

Name	Event Categories:	Performers	Foreground Situation	Background Situation
Weizmann [Gorelick *et al.* 2007]	Wave/Walk/Run/Bend/Skip/Side/Jump	9	Single human Clear Boundaries	Solid Colour
KTH [Schuldt *et al.* 2004]	Wave/Walk/Run/Boxing/	25	Single human Multiple Viewpoint Clear Boundaries	Solid Colour
Campus	Wave/Walk/Run/Bend	3	Single human Complex textures	Complex textures and moving objects

Table 6-1 Selected datasets used for evaluations

Figure 6-1 Snapshots from Weizmann datasets

Since the clips from the Weizmann video library contain mostly static and relatively static simple backgrounds with only one actor in each file, it is considered an ideal source for generating event templates in this project. The datasets provide clear human contours, taken from a fixed video camera position with fixed internal parameters that can be readily used to define human actions models using Active Contour segmentation techniques.

Figure 6-2 Snapshots from KTH datasets

In contrast, the KTH dataset contains more than 25 event performers and variant actor scales and camera positions that are ideal to test the devised STV event processing techniques and algorithms against other existing approaches using the same dataset. Further variations have been added into the experiments, for example, the illuminations of the KTH video clips had been changed to ±20% in each test to analyse the process stability, which guarantees sufficient light for human vision system and also avoids unnecessary glared areas hampering the recognition operations.

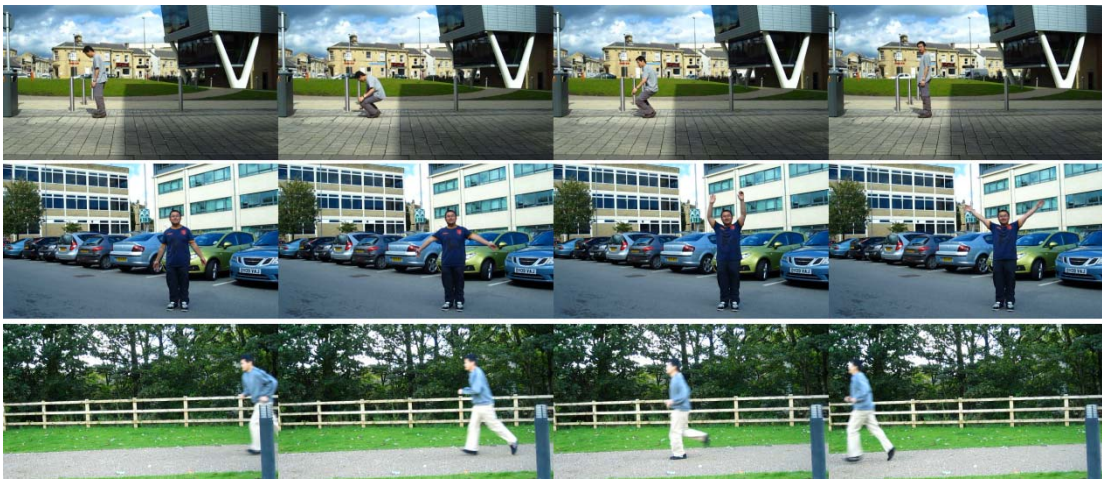


Figure 6-3 Snapshots from the self-made Campus datasets

To evaluate the proposed STV-based event detection theories and the prototype's robustness performance under real application conditions, a large set of footages have been recorded in this programme at the University of Huddersfield campus. Selected snapshots are shown in Figure 6-3. This dataset is focused on system robustness performance under complex background and illumination changes. These challenging backgrounds both contain large blocks of uniform colour regions and many small textured areas from static and moving objects similar to typical real-world dynamic noises CV tasks are facing.

6.2. System Performance Evaluations

6.2.1. Evaluation Benchmark

In typical pattern recognition systems, especially the binary decision making systems, the recognition accuracy is often evaluated by two statistical figures: false positive and false negative rate. The false positive rate, on one aspect, highlights how many actual negative samples are treated as positive ones. For, example, in a "waving" event detection system, a false positive case means the background noise is mistaken to a human "waving" event. The false negative instance, on the contract, means positive cases being overlooked and are treated as negative samples. The relationship between false positive and false negative can be illustrated in the so-called confusion matrix as shown in Table 6-2. (T, F, P, N are the abbreviations of True, False, Positive and Negative, respectively).

	Actual Positive	Actual Negative
Predict Positive	TP	FP
Predict Negative	FN	TN

Table 6-2 Confusion matrix

In the experiment and evaluation part of this thesis, the performance of the system accuracy is evaluated by changing the prediction threshold from 100% to 0% to form the "Receiver Operator Characteristic curves (ROC curves)" and "Precision Recall Curve (PRC). The definition of those curves can be explained by following equations, more details can be found from in Davis and Goadrich's paper [2006].

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{True Positive Rate} = \frac{TP}{TP + FN} = \text{recall} \qquad \text{False Positive Rate} = \frac{FP}{FP + TN}$$

6.2.2. Efficiency Measurements

The overall time consumption of the devised STV event matching algorithms is composed of three parts: STV construction, model segmentation, and RI-matching. Based on the preliminary trials, the main deciding factor for the system efficiency is the total number of selected sub-regions accumulated at segmentation stage. Hence, the efficiency evaluation designed in the experiments has largely focused on this part as well as RI-based matching, which takes the assembled sub-regions as the input. As illustrated in Figure 6-4(a), a mock event template is defined as a volumetric sphere. Therefore the contributing-to-non-contributing voxel ratio is controlled by the sphere size within the model. Experiments have been carried out to establish the relationships between a template's size inherent and its time consumptions at the template matching stage. Figure 6-4(b) shows a 512×512×512 STV block that can be considered as a volumetric model formed by the same number of small cubical "bounding" boxes organised in an octree style. The total number of sub-regions n need that to be tested for intersections in the I-PWRC pipeline are defined as 8^0, 8^1, 8^2 and 8^3 - representing different hierarchical levels of segmentation. This simplified design provides an indicator on the relationships between the STV segmentation sizes and their matching

speeds. To further simplify the benchmark design, the position of the mocking event template has always been fixed at the centre of the STV.

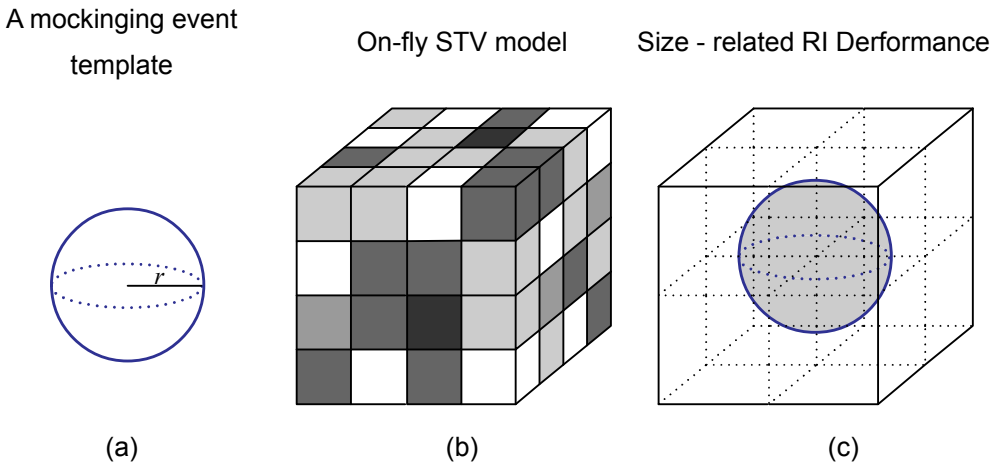


Figure 6-4 Artificial event model and STV hierachical strcture for efficiency evaluations

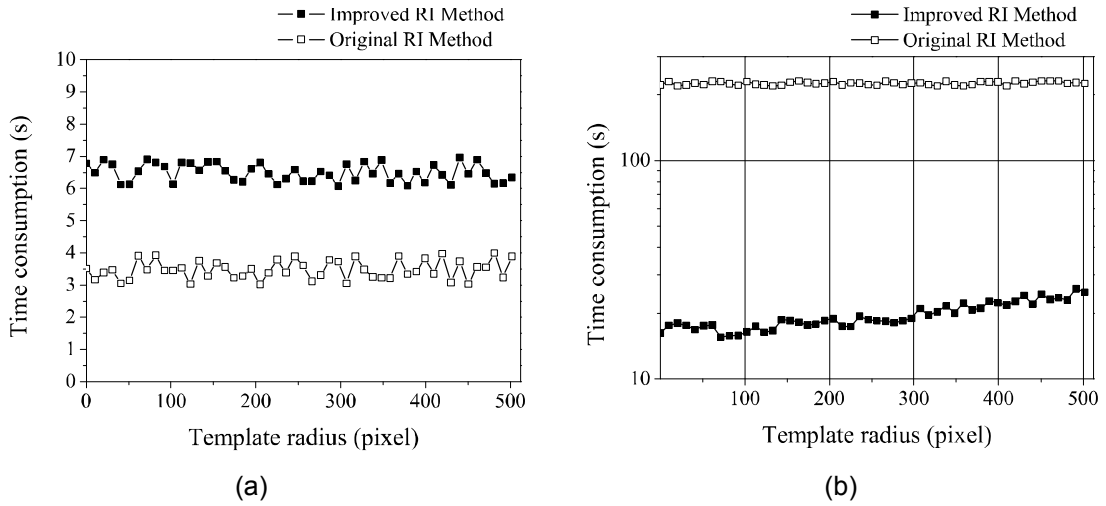


Figure 6-5 Improvements on time consumptions from the "Filtering" and "RI Matching" phases

Figure 6-5 demonstrates the time consumptions of the original RI method and the improved one developed in this research at different segmentation levels. As indicated in Figure 6-5(a), at the segmentation level $n=8^1$ (level 1 division in an Octree), the original method outperforms the proposed method by a fold due to the latter's extra filtering operations. However, when the segmentation level increases to 8^3 as shown in Figure 6-5(b), the new approach has displayed a 10-plus efficiency gain. Based on the experiments, the original RI method runs faster if the size of each sub-region is

not “too small” compared with the template; while the improved RI method produces a far superior performance under real application conditions when complex and dynamic events have to be extracted through over-segmentation to preserve details.

The event template size used for RI matching is also a contributing factor to the efficiency of the devised algorithms. When an inputting STV is extensively over-segmented to handle noisy signals, the increase of a template’s size can bring extra cost to the operational time but still only counting as a fractional cost of the entire processing time.

6.2.3. Matching Accuracy Evaluations

Experiments were carried out in this research to assess the event detection accuracy based on the theoretical structure of the system as introduced in Chapter 3 and 4.

One of the main objectives of these tests was to establish the ground truths on event detection accuracy of the proposed method. The experiments started from the KTH video libraries. Table 6-3 lists the values of the parameters used in the experiments, where the event templates were defined by averaging the minimum of four volumetric contours extracted from each event category.

MS		PWRC		Active Contour		Coefficient Factor	
H_r	H_s	$k(C_0)$	n	α	β	a	b
5	5	0.2	7	0.14	1.17	0.8	0.6

Table 6-3 Parameters used for KTH Dataset

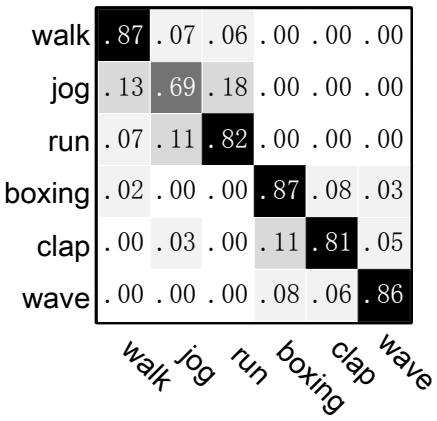


Figure 6-6 The KTH confusion matrix

Figure 6-6 shows the test results of the detection accuracy based on the confusion matrix acquired from the KTH dataset. The average accuracy of the developed system is 82.0%, which is slightly better than many popular methods as listed in Table 6-4. As illustrated in the confusion matrix, certain events such as the jog-and-run and the boxing-and-clap pairs are difficult to distinguish due to their silhouette similarities and minute variations on the temporal axis. One possible solution to such a problem is to combine the machine learning algorithms with the local spatio-temporal features for differentiating the details of human gestures.

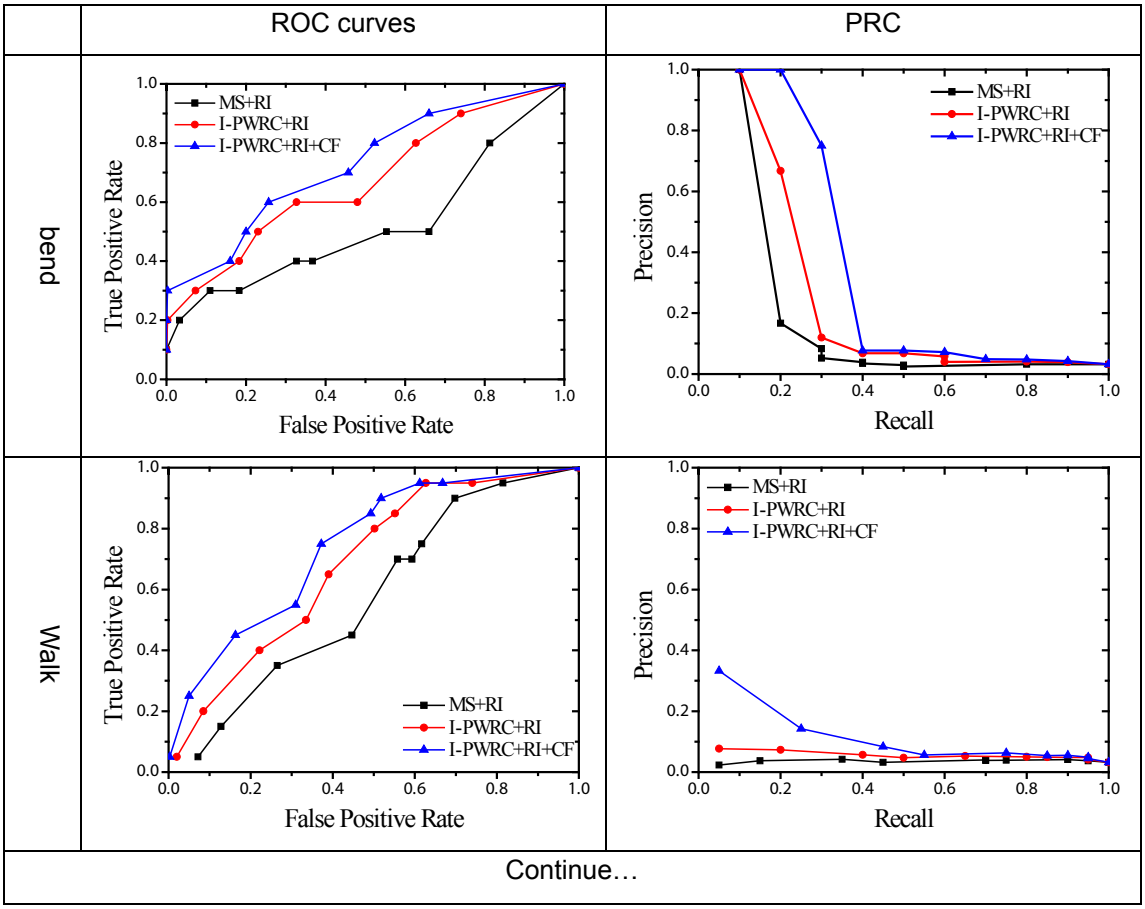
Methods and techniques	Event Detection Accuracy
This Method: PWRC + RI + CF(Coefficient factor)	**82.0%**
Ke *et al*.'s MS (Mean shift) + RI + Flow [2010]	80.9%
Schuldt *et al*. [2004]	71.7%
Dollár *et al*. [2005]	81.2%
Niebles *et al*. [2008]	81.5%

Table 6-4 Matching accuracy performance compared with other approaches

The self-made Campus datasets have also been tested, which contain mixed action events recorded under different uncontrolled and real-world conditions. The length of

each video clip is about 10 minutes. The prototype system's robustness has then been validated by using the ROC curves and PRC as illustrated in Figure 6-7.

Figure 6-7 presents the ROC curves generated from relevant experiments to highlight the performance variations on the proposed event detection algorithms recorded at an incremental the threshold value (+10% for each plot in the curves). It is evident from the results that the proposed method can produce a better performance through integrating the coefficient factor mechanism as explain in Section 4.3.2. In addition, the innovative Hierarchical I-PWRC technique can abstract more accurate shape features in comparison with other clustering based segmentation methods.



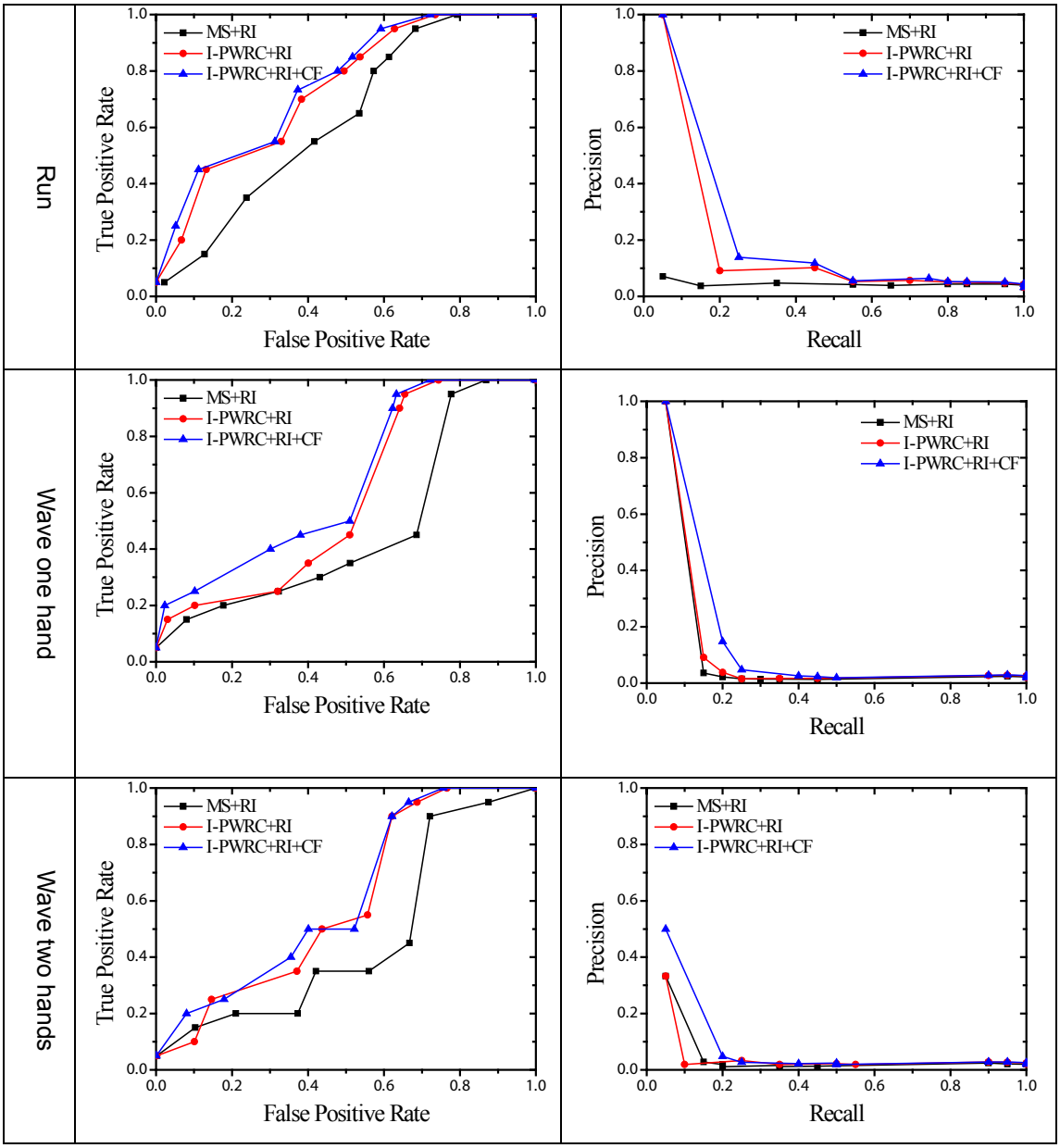


Figure 6-7 Template matching results (ROC and RP curves) on the campus dataset

6.2.4. Model Compatibility Evaluations

As discussed in relevant sections, various new algorithms have been introduced into the process pipeline. Their integration smoothness will have a major impact on the overall system performance. This experiment aims at assessing the system module compatibilities, especially between the data preparation (MS pre-clustering) and the I-PWRC segmentation; and the RI matching operations. The normalised RI distance

introduced in Section 4.2.2 (see Equation 4-4) has been used as the main indicator for the compatibility performance.

The STV event templates deployed in this experiment have been drawn from the Weizmann's datasets with the test videos coming from the Campus' dataset. As extensively covered in Section 3.2, both the STV preparation and the region graph construction stages have employed segmentation operations.

As illustrated in Figure 6-8, smaller over-segmented regions produced by MS ($h_c < 15, h_l < 15$) are often used in the conventional RI matching applications which contain finite region boundary sections, but are often too small to be assembled to represent the event pattern shapes effectively. Although this problem can be partially relieved by using the normalised distance and the coefficient factor during RI matching, a more fundamental solution still has to come from the accurate segmentation outputs.

As shown in Figure 6-9, larger MS-driven over-segmented sub-regions ($15 \le h_c \le 30$ and $15 \le h_l \le 30$) encapsulate uniform coloured regions well but are easily confused by similar colours from separate objects and missed out on some small textured areas completely, especially in the low illumination and contrast areas. The balance between choosing larger or smaller segmented regions can be a tricky one and is only controlled by the MS clustering window size h The time consuming job on adjusting this sensitive parameter can reduce accuracy, adaptability and robustness of the overall system. This effect can be illustrated by the confusion matrix shown in Figure 6-10, which compares the detection accuracy using conventional RI method (based on MS) and the I-PWRC segmentation method.

Figure 6-8 MS over-segmentation result (h_c=5, h_l=5) on 3 events

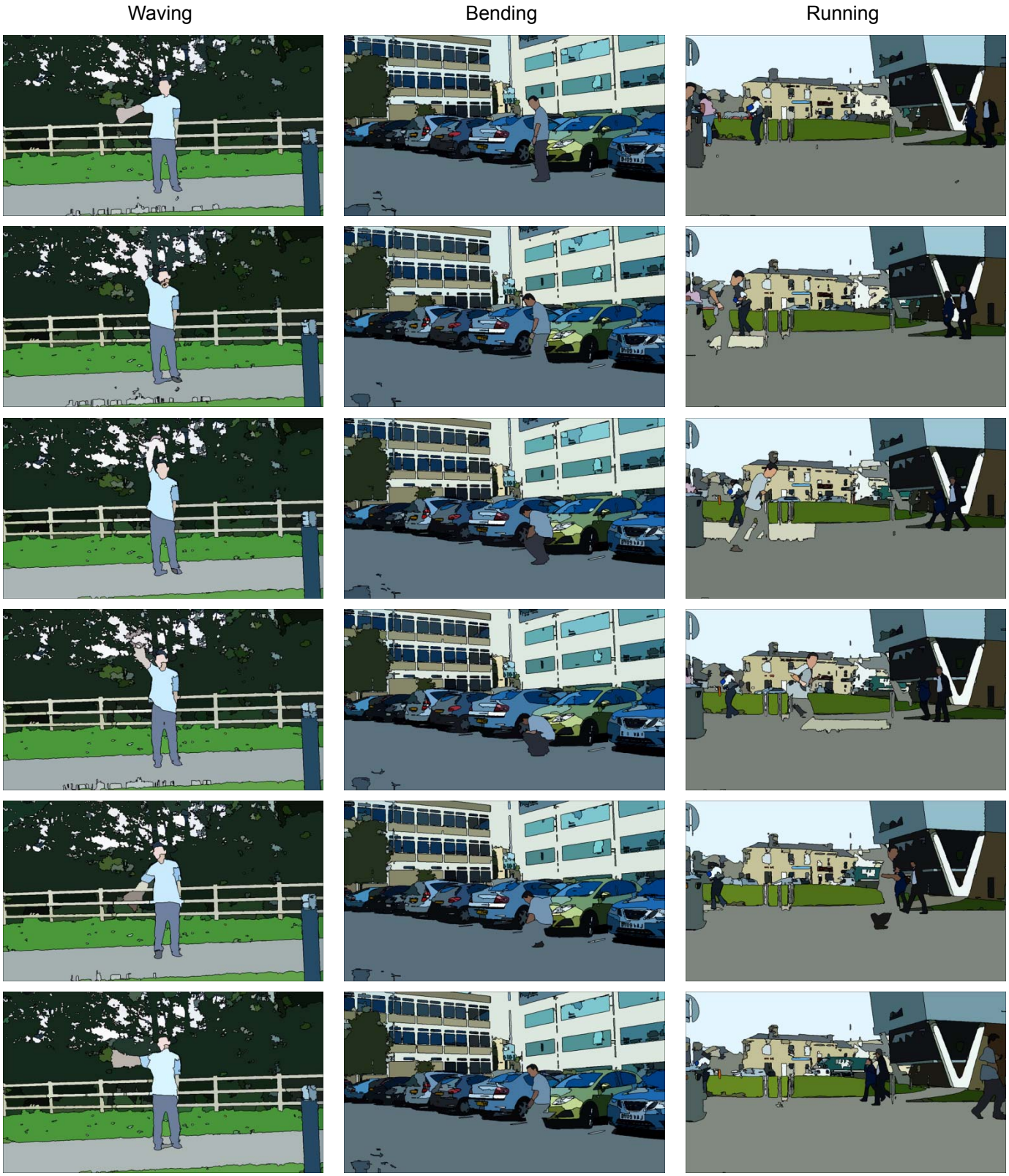
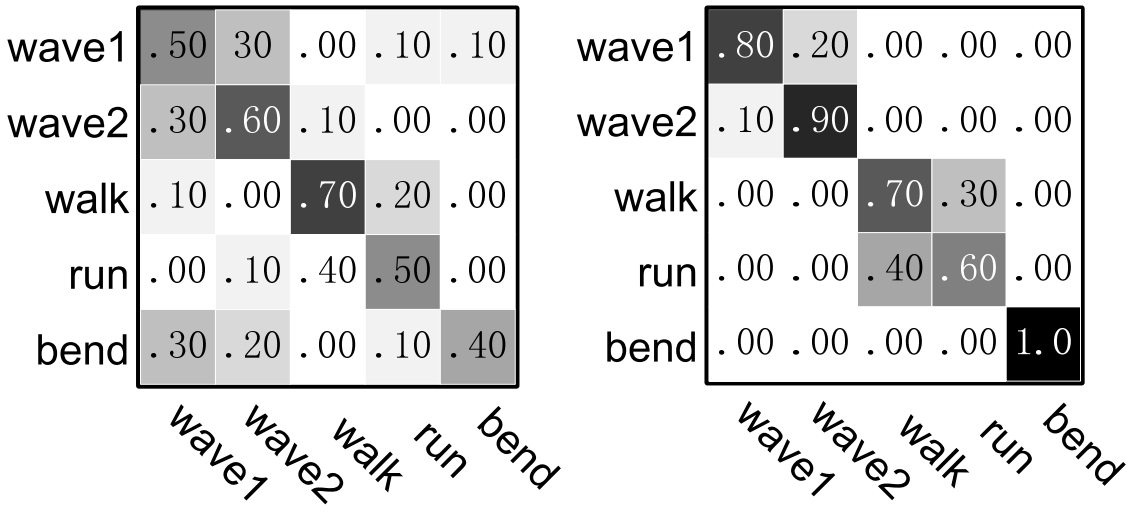


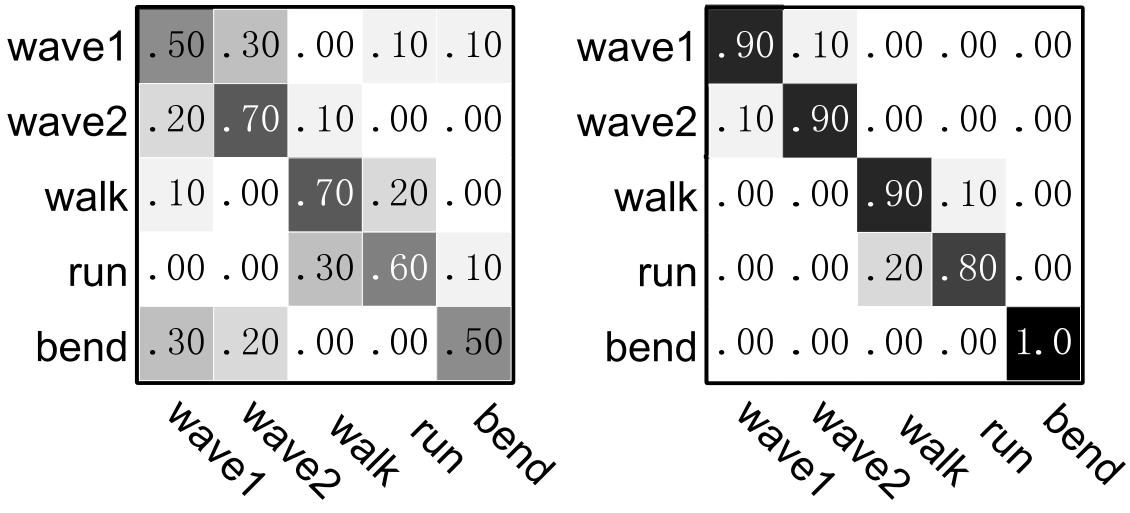
Figure 6-9 MS segmentation results (h_c=15, h_l=15) on 3 events

It is evident in the experiment that the I-PWRC method has been benefitted greatly from its hierarchical structure and the dynamic parameter controls. The accuracy performance, therefore, is much better than the benchmarked MS-only approaches. Similar conclusion can be reached by comparing the I-PWRC-empowered method with Coefficient Factor (CF)-boosted RI method as shown in Figure 6-10(c) and Figure 6-10(d). The event detection accuracy has improved by more than 9% by using I-PWRC.



(a) MS+RI, average accuracy = 57.1%

(b) I-PWRC+RI, average accuracy = 80.2%



(c) MS+RI+CF average accuracy = 61.1%

(d) I-PWRC+RI+CF average accuracy = 90.1%

Figure 6-10 MS and I-PWRC based RI matching confusion matrices

Based on the experimental results, the I-PWRC method has proven its effectiveness for providing quality outputs for the RI matching operations, especially under complex real world conditions. Accompanied by the optimisation measures employed at the template matching stages, the process pipeline and its various operational models have shown sound compatibility.

6.2.5. Matching Performance within Different Colour Spaces

Similar to the gray-scale-based intensity features in image processing, colour features are becoming more popular in modern DIP and video processing backed up by many new segmentation and matching algorithms as detailed in Chapter 2. This experiment has focused on the devised algorithm performances in different colour spaces, which highlight the system accuracy improvement contributed by low-level features.

The datasets used in this experiment are the same as the ones deployed in Section 6.2.3. As illustrated in Figure 6-11, the average accuracies calculated from the confusion matrices based on the I-PWRC segmentation outputs are very close and maximum differences are within 1%.

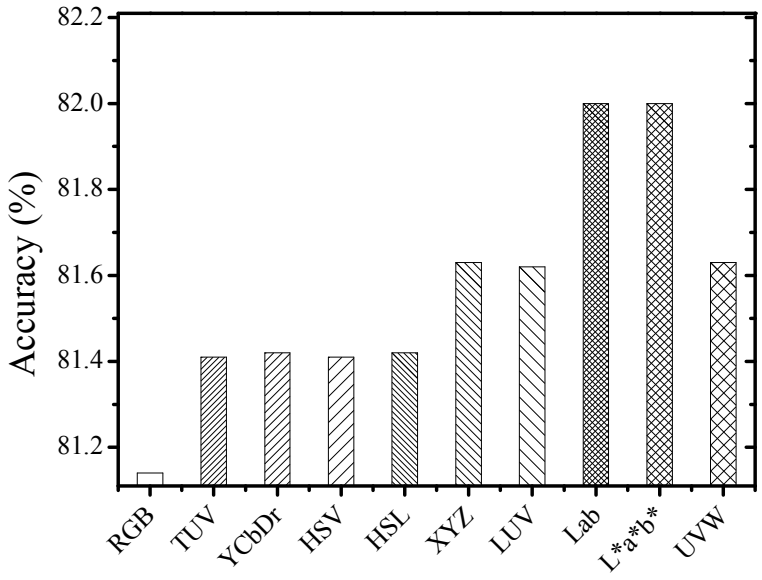


Figure 6-11 KTH datasets average detection accuracy based on different colour spaces

Overall, the group of CIE colours (XYZ, LUV, Lab, L*a*b* and UVW) performed better than others due to their human perception-based colour representations. In addition, the Luminance plus Chrominance group (TUV and YCbDr) have shown identical detection accuracies comparing with the Hue and Saturation (HSV, HSL) group.

As a conclusion, although most of today's commercial video imaging sensors and equipment still adopt RGB-based colour settings, for automating video analysis and event detection tasks in the future, more human perception-based and computationally efficient colour models should have been deployed.

6.2.6. Template representation and Matching Performance

In the system and experiment design, the event templates are composed by averaging a number of event models extracted from each event clip group. The purpose of this experiment is to establish the relationships between the matching accuracy and the template representativeness denoted by the number of samples used for generating a

template. The event samples were selected from the Weizmann dataset and the test videos were coming from the Campus video dataset. The matching accuracy performances of each event category are illustrated in Figure 6-12.

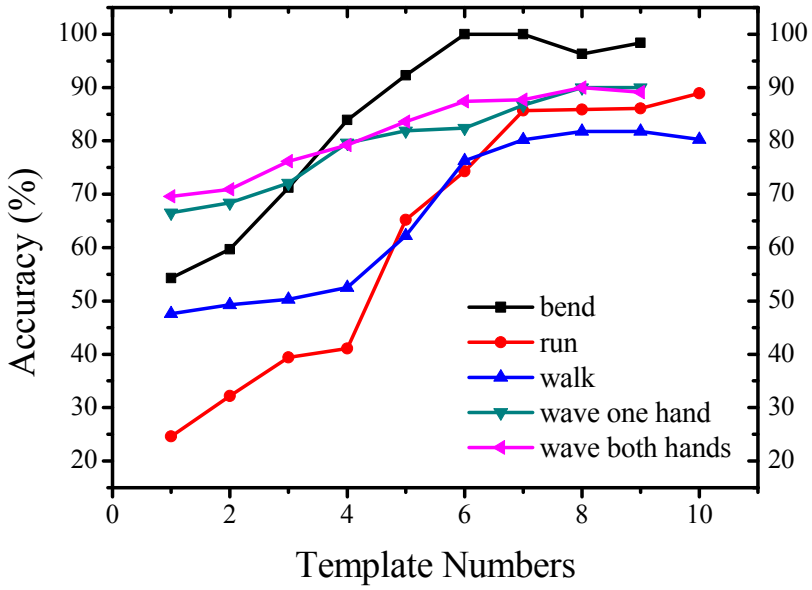
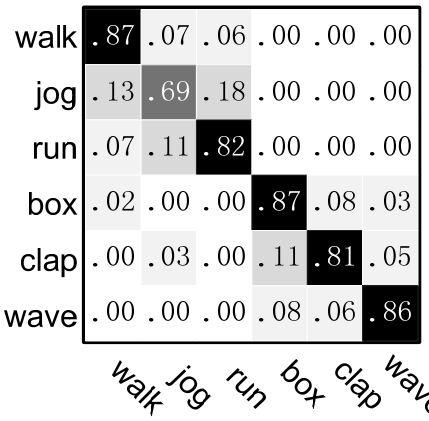


Figure 6-12 Accuracy impact of average template numbers

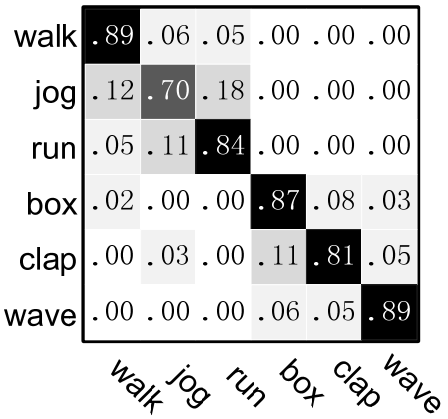
It is evident in the figure that the matching accuracies for all event categories peaked around the sample size of 7. Further increasing the sample size produces small variations on matching accuracy with the event groups such as “walking” and “bending” even shows a small drop on accuracy due to unnecessary sample details.

6.3. Scale-Invariant Event Detection

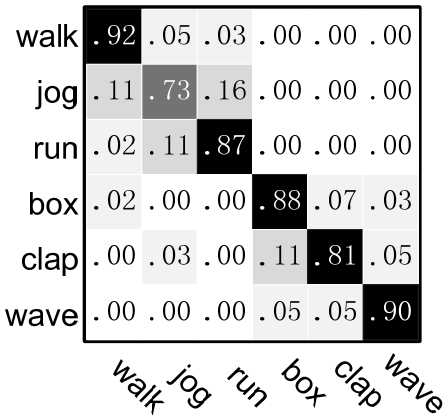
Figure 6-13 (a), (b) and (c) provided direct comparisons between a benchmarking confusion matrices generated from the KTH dataset using a standard matching approach and the multi-scaled templates introduced in Section 5.1.



(a) Benchmarking detection results (without robustness measures and average accuracy at 82.0%)



(b) Improved by multi-scaled templates (average accuracy=83.4%)



(c) Improved by normalised multi-scaled templates (average accuracy=85.1%)

Figure 6-13 KTH confusion matrices after employing the using multi-scaled templates

Through examining Figure 6-13 (a) and (b), it is clear that the multi-scaled templates have introduced consistent accuracy improvements on all tested KTH samples which contain large variations on event spatio and temporal features. The "false-negative" incidents, such as the miss-identification of run from walk have been reduced significantly through employing the multi-scale templates. In addition, the inherent "double check" mechanism from using the multi-scaled templates has also improved the "true-positive" rate through cutting the "false-positive" parts. By using the multi-scaled templates, the tested dataset has seen an average improvement at 1.4%. As shown in the Figure 6-13(c), the accuracy further improved to 3.1% by using the

normalised version of the multi-scaled templates. The test results on the Campus datasets are illustrated in Figure 6-14.

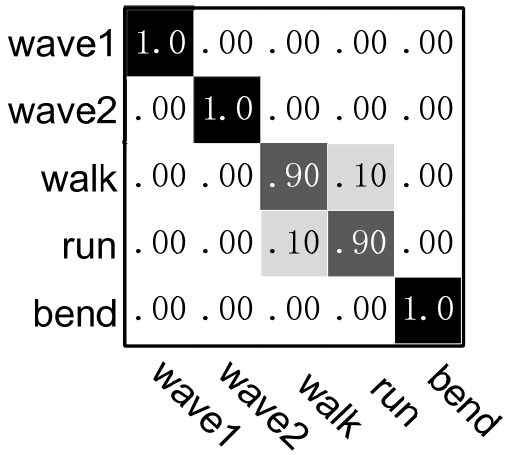


Figure 6-14 Normalised multi-scaled templates applied on the Campus dataset

6.4. Test on Uncontrolled Video Inputs

Partially serving as a proof-of-concept as well as performance evaluations on system performance, the afore-discussed experiment results have proven the validity of the devised STV-based event detection approach and the practicality of its corresponding system design. The experiments introduced in this section have been focusing on the real-world system performance when subjected to random noise and other challenging real application conditions such as dynamic background and illumination changes.

Figure 6-15 Selected frames from a "people falling down" video clip

The real-world CCTV video files were downloaded from various public websites. Figure 6-15 shows the snapshots of a number of pedestrians tripping and falling down on a spot near the entrance of a building. The CCTV postures are controlled by a rotational motor with adjustable focal length. The experiment intended to detect and denote all the "falling down" events from the video footage that lasted for about 8 minutes-there are 12 actual "falling down" events captured on the tape confirmed by visual observations. It is clearly visible from the selected snapshots, the background of the video was filled with noise signal, i.e. moving vehicles and passing pedestrians. The "falling down" event template is generated from the test video itself.

Figure 6-16 Over-segmented sub-regions from the "falling down" event

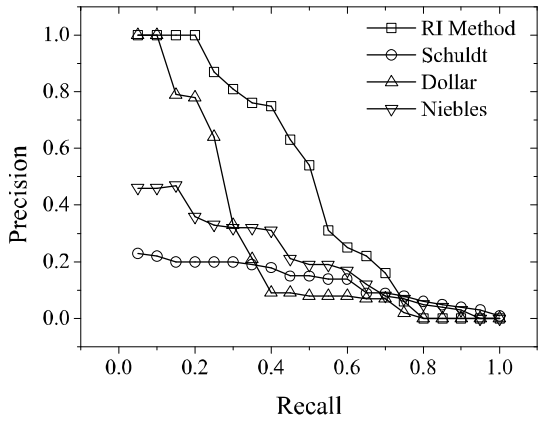


Figure 6-17 Performance comparisons between classic approaches

Figure 6-16 shows the I-PWRC segmentation process applied on the input video. Once the 3D over-segmentation regions are generated, the template is then being used to "scan" through the entire spatio-temporal domain for matching operations. Figure 6-17 compares the RPC drawn from a number of classic matching algorithms (see Table 6-4), which shows a slightly superior performance from the improved RI method. The experiment had set the threshold at 60% for a positive match in between the pre-defined event template and the on the fly STV models.

Chapter 7. Conclusions and Future Work

7.1. Programme Summaries

This research programme has been focusing on tackling problems in video event detection based on the so-called Spatio-temporal Volume and its related voxel-based features. In this research, video contents have been modelled as volumetric shapes for event recognition through devising innovative feature extraction and shape matching techniques.

7.1.1. I-PWRC Validity and Practicality

In this dissertation, STV feature extraction problems are identified and investigated by harnessing the advancement and potential of classic image segmentation and pattern recognition techniques. An innovative image segmentation technique has been developed in this research during the feature extraction phase of the operational pipeline. The proposed extended Pair-wise Region Comparison (I-PWRC) method established a set of hierarchical segmentation operations for classifying STV regions based on regional colour and texture features. Its baseline algorithm follows an iterative mechanism and updates each cluster in every cycle by comparing their inner difference and similarities with neighbouring clusters. In this research, a graph-oriented comparison approach has been successfully implemented into the STV space.

Based on the theoretical study and practical trials, the I-PWRC segmentation strategy developed in this project has proven its effectiveness and efficiency when extended from 2D to 3D feature spaces. The 3D segmentation approach is not merely an

extension of the feature vectors from their original forms, but a methodology remodelling through analysing STV feature characters and their spares matrix nature. The system prototype developed in the project further verified the innovative approach's practicality through series tests on volumetric models and their huge amount of data. This research has opened up valuable algorithm design approach when dealing with problems rising from data structural and scale changes.

Currently, the system pipeline employed a separate MS pre-clustering operation for reducing the complexity of the I-PWRC graph initialisation, the performance of the system has therefore been partially depending on the output qualities of the MS clustering process. The additional parameters engaged for maintaining the MS performance on different datasets increase the variety of the overall segmentation performance due to the empirical maintenance during the feature extraction steps. It is anticipated that this problem can be tackled by integrating the MS-based pre-clustering and I-PWRC segmentation into one unified operation through harnessing the recent advancements in the so-called sparse feature representation.

In this research, the developed texture-based graph feature representation method has been proven as an effective approach for 3D feature segmentation that is a significantly improvement from the conventional per-frame and pixel level operations. The I-PWRC produces a global STV feature representation, while the texture and its histogram representation enhance the segmentation accuracy through applying local features to the global representation. This hybrid approach had been stimulated by the classic success in face recognition when combining the local and global features such as the Eigen-face method. One of the drawbacks for the texture-based representation rooted from its computational cost that hampered its applications. This problem will

be ideally dealt with by the computer hardware and programming paradigm shifts, such as Cell-CPU, multi-core and Graphic Process Unit (GPU)-driven hardware acceleration and parallel computing.

7.1.2. 3D RI Matching Adaptability and Robustness

Based on the early works from Ke *et al*.'s[2010], an improved Region Intersection (RI) method has been developed for recognising video events by comparing the global features assembled from the over-segmented I-PWRC outputs with the event templates. The matching outputs are then further refined by deploying an evaluation scheme called coefficient factors to assess the matching (RI) distances. These devised recognition procedures have shown their distinctive advantages when dealing with real-world video inputs containing dynamic background and noisy signals. This research has also introduced the scale-invariant templates for matching calibration. Based on the test results, these developments can improve the robustness performance of the RI matching in uncontrolled videoing conditions, especially when the video inputs contain camera posture transformations.

It is worth noting that histogram representation and its distancing measurement have been applied in the coefficient factor phase for boosting the matching performance. Similar histogram-based methods have also been successfully deployed for representing textures in the I-PWRC segmentation stage in the research pipeline, as in many other successful vision applications, such as [Dalal and Trigges 2005], [Wang and Mori 2009] and [Grundmann 2010]. The usage of histograms therefore can be recognised as an important tool for bridging the gaps between low-level and high-level feature representation methods, and individual feature points and feature groups in different feature spaces.

Another interesting finding from this research is that the 3D RI matching approach is a adaptable global feature-based technique for comparing the feature distributions from over-segmented STV regions with the pre-defined 3D event templates. While many recent publications have focused on local STV features such as spatial body part relationships (see Section 2.5), the evaluation in this research has shown superior and/or comparable detection accuracy under identical video settings using the devised "global" feature-oriented methods. The rationale of this phenomenon can be summarised as: firstly, the STV feature space provides natural and comprehensive information for modelling video content and their dynamics defined as over-segmented feature regions. This intuitive global representation simplifies the matching process into a so called "in-class variations" operation, which is ideal for maintaining system robustness in comparison to complex local feature-based methods. Secondly, since most local feature-based event detection techniques require machine learning strategies for establishing event categories and matching rules for a particular event, the computational expensive deductive reasoning procedures are not suitable for the inherently large-scale STV models. The 3D RI matching theory, only based on Set Theory, is an effective method for the STV template matching.

As explained in Section 4.4, the over-segmentation design currently employed in the research cannot comprehensively represent event information which is "inside" a STV model (i.e. a concave shape). The global representation strategy is seemingly lacking in intrinsic characteristics in dealing with video occlusion problems, which justifies the motivation in this research to apply a hybrid (global and local features) mechanism for the problem. Other envisaged optimisation strategies have been discussed in Chapter 5 and classified as follow-on works of this programme.

7.1.3. Function Modularisation and System Integration

To maintain runtime performance of the research system, the process pipeline within the system prototype adopted a modularised design and has been enhanced by a number of optimisation techniques, for example, the volume buffering mechanism for storing STV data through composing models based on the incoming video streams in a compact and on the fly style to control the runtime memory consumption. The research prototype has also demonstrated the innovative interest area-based (not interest point) data structures for improving computational efficiency of the system. The interest areas predict and highlight the likely event-occurring areas before deploying the sliding window filtering mechanisms for shape matching using the SIFT feature points. Experiments carried out in the project have justified the feasibility and flexibility of these optimisation designs.

In the experiments, (see Section 6.4), under challenging videoing conditions, a combination of those measures in addition to the scale-invariant templates mechanism have ensured a satisfactory overall system performance against other benchmarking methods and systems.

Although there are still many issues to be studied and resolved, the STV-based video event detection strategy and the related techniques developed in this research have revealed the validity and potential of the approach for tackling the challenging problems rooted from real-world video processing and semantic interpretation for a wide spectrum of applications.

7.2. Future Work

The STV model and its related algorithms investigated in this research have justified their values in tackling the challenging problem of video-based event detection, which is considered a small step towards the ultimate solution for real-time intelligent and automated complex event detection and recognition. Although this "Holy Grail" in the CV field cannot be readily achieved just using current knowledge and techniques, many worthwhile attempts aimed at that goal had offered remarkable ideas and initiatives that will facilitate future efforts. This section covers a broad discussion on several related research and development directions for future exploration.

- Compressed 3D Video Feature

Current STV-based volume data techniques are mainly based on extending 2D DIP algorithms into 3D domains. However, this approach often suffers from the so-called "curse-of-dimensionality" due to the complex model structures and feature definitions introduced. For example, to apply classic PWRC techniques directly into a 3D feature domain, the initial 8× pixel linkages for each feature point assessed will be expanded to 26× potential voxel connections. Considering the iterative processing natures of many relevant operations, the computational burden generated from such a shift can be daunting and even impractical at time. In this research many remedial measures have been taken to address the issue, such as MS pre-segmentation and region histogram representation, which observed relative success to the proposed methodology. On the other hand, the full potentials of 3D PWRC were not entirely realised due to the quality of segmentation being partially affected by the outputs of the pre-processing steps.

One possible solution to this problem is to renovate the conventional method for video STV construction. Motivated by the rapidly evolving video compressing techniques, the design principle is to move away from the current STV data structure that relies on uncompressed video frames, and can only be processed by voxel-based algorithms, to a new paradigm based on compressed video formats and to integrate relevant video compressing algorithms (codec) with appropriate feature processing techniques. It is anticipated that efforts along this direction will transform the spatial feature dominated STV processing into a frequency feature analysis domain, hence, enabling many powerful and mature analytical models and tools to be used.

- Comprehensive Local and Global Hybridisation

In a typical computer vision application, the "meaningful" information is fundamentally represented by pre-defined features, which determine the appropriate analytical methodologies in the following processing steps. Based on the uncompressed low-level voxel characteristics, such as colour, intensity and spatial positions, the features used for event detection in this research were predominantly based on 3D shapes, regions textures, and sudden colour/intensity changes (interest feature points). Determined by the inherent nature of those features, the research problems have been tackled by investigating specific global segmentation and template matching techniques. However, as discussed in relevant sections, local feature-based analysis can still provide substantial benefits on result verification and performance enhancement. For example, the popular Active Contour method for tracking might rely on manual selection of event targets at the start. But its inherent energy minimisation procedures will enable the rest of the processes to be automated. Another classic example is the so-called "Optical Flow" that had been a research

hotspot for almost 2 decades. It has been widely used in motion estimation from multiple and continuous frames. It is envisaged that to combine the strength of voxel-based and "flow"-based features, the "concave" event shape identification problem can be alleviated. If facilitated by appropriate machine learning algorithms, the entire process might even be integrated and automated for practical applications.

- New Sensor Technologies

Benefited from the advancements of sensor technologies, such as Charge-Coupled Devices (CCDs) / Complementary Metal–Oxide–Semiconductor (CMOS) image sensors and depth sensors, many legacy software/algorithm-driven calculations have been moved to hardware paradigms and being directly "measured" and recorded, which might lead to a revolution for future video and CCTV applications.

In 2004, the first off-the-shelf time-of-flight (ToF) camera has been released by Advanced Scientific Concepts (R), Inc. [2011]. Combined with traditional CCD sensor technology, ToF introduced an extra depth sensor that measures the distance between a target object and the lens before producing a 3D depth map of the captured scene. Analogical to the radar system, ToF measures the temporal duration of light leaving and reflected back to the camera, to register range information with an effective distance up to 60m, and a resolution of about 1cm. Different from conventional 3D reconstruction techniques, this added dimension enables 3D reconstruction from a single camera instead of two or more cameras working with time-consuming 3D reconstruction algorithms. The technology has been successfully applied in modern video game designs. To apply the ToF in video event detection should reduce the burden being carried by most of today's application systems in

terms of background removal, human body segmentation, and occlusion detection. It should also provide accurate object contours for image-based rendering.

Benefiting from current hardware developments in image sensors, some features such as speed and depth, which cannot easily be abstracted from video cameras, can boost the performance of video event detection significantly.

- Computational Hardware Acceleration

The RI-based template matching in the research system is considered as the most complicated operation counting at 70% of total CPU time consumption. Although this project was not initially targeting real-time applications, the operational efficiency still plays a vital role for the proposed method's future success and wider applications. One perspective solution to improve the system efficiency is through employing hardware acceleration by adopting parallel computing architectures, for example, through harnessing data parallelism embedded in modern Graphics Processing Units (GPUs) to facilitate inherent data intensive and filter-driven video feature computations. The Compute Unified Device Architecture (CUDA) developing platform from nVIDIA [2011] has provided implementation tools for this purpose.

GPU is skilled in processing a same algorithm on large quantity of data, many important steps in the research pipeline can be accelerated by using this device. For example, the interesting feature points of STV can be extracted by translating the voxel data as 3D textures, which can be processed by fragment shading of GPU. The following machine learning steps can also gain great benefit by defining the feature points as texture (2D or 3D), which can be deployed by various shading languages as texture maps during processing. These improvements provide a possible real-time solution through accelerating the volume-based operations developed in this research.

References

Adams, R. and L. Bischof (1994). "Seeded Region Growing." IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(6): 641-647.

Adelson, E. H. and J. R. Bergen (1985). "Spatiotemporal Energy Models for the Perception of Motion." Journal of the Optical Society of America A **2**(2): 284-299.

Advanced Scientific Concepts, I. (2011). Advanced Scientific Concepts, Inc. ASC 3D Portable camera shipping since 2004.

Agarwal, S., A. Awan, et al. (2004). "Learning to Detect Objects in Images via a Sparse, Part-based Representation." Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(11): 1475-1490.

Agarwala, A., M. Agrawala, et al. (2006). "Photographing Long Scenes with Multi-viewpoint Panoramas." ACM Transactions on Graphics **25**(3): 853-861.

Ahmed, M. N., S. M. Yamany, et al. (2002). "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data." Medical Imaging, IEEE Transactions on **21**(3): 193-199.

Alper, Y. and S. Mubarak (2005). Actions Sketch: A Novel Action Representation. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Baker, H. H. and R. C. Bolles (1989). "Generalizing Epipolar-Plane Image Analysis on the spatiotemporal surface." International Journal of Computer Vision **3**(1): 33-49.

Baker, S., D. Scharstein, et al. (2007). A Database and Evaluation Methodology for Optical Flow. IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.

Beauchemin, S. S. and J. L. Barron (1995). "The Computation of Optical Flow." ACM Comput. Surv. **27**(3): 433-466.

Bobick, A. F. and J. W. Davis (2001). "The Recognition of Human Movement using Temporal templates." Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(3): 257-267.

Boden, M. (2006). Mind as Machine: A History of Cognitive Science. Oxford, England, Clarendon Press.

Bommer, J. J., F. Scherbaum, et al. (2005). "On the Use of Logic Trees for Ground-Motion Prediction Equations in Seismic-Hazard Analysis." Bulletin of the Seismological Society of America **95**(2): 377-389.

Bregonzio, M., J. Li, et al. (2010). Discriminative Topics Modelling for Action Feature Selection and Recognition. British Machine Vision Conference 2010. F. Labrosse, R. Zwiggelaar, Y. Liu and B. Tiddeman. Aberystwyth, UK, BMVA Press**:** 8.1-8.11.

Cha, S. H. and S. N. Srihari (2002). "On measuring the distance between histograms." Pattern Recognition **35**(6): 1355-1370.

Coifman, B., D. Beymer, et al. (1998). "A real-time computer vision system for vehicle tracking and traffic surveillance." Transportation Research Part C: Emerging Technologies **6**(4): 271-288.

Comaniciu, D. (2002). "Mean Shift: A Robust Approach Toward Feature Space Analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**: 603-619.

Cui, M., J. Femiani, et al. (2009). "Curve Matching for Open 2D Curves." Pattern Recognition Letters **30**(1): 1-10.

Dalal, N. and B. Trigges (2005). Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005. San Diego, USA. **1:** 886-893.

Datta, R., D. Joshi, et al. (2008). "Image retrieval: Ideas, Influences, and Trends of the New Age." ACM Computing Surveys **40**(2): 1-60.

Davis, J. and M. Goadrich (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, ACM**:** 233-240.

Davison, A. J. (2005). Active search for real-time vision. Tenth IEEE International Conference on Computer Vision, ICCV 2005. .

Dollár, P., V. Rabaud, et al. (2005). Behavior recognition via sparse spatio-temporal features. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on**:** 65-72.

Duda, R. O. and P. E. Hart (1972). "Use of the Hough Transformation to Detect Lines and Curves in Pictures." Communications of the ACM **15**(1): 11-15.

Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. Hoboken, Wiley.

Due Trier, O., A. K. Jain, et al. (1996). "Feature extraction methods for character recognition-A survey." Pattern Recognition **29**(4): 641-662.

Fathi, A., M. F. Balcan, et al. (2011). Combining Self Training and Active Learning for Video Segmentation. British Machine Vision Conference 2011.

Feddema, J. T., C. S. G. Lee, et al. (1989). Automatic Selection of Image Features for Visual Servoing of a Robot Manipulator. IEEE International Conference on Robotics and Automation. **2:** 832-837.

Felzenszwalb, P. F. and D. P. Huttenlocher (2004). "Efficient Graph-Based Image Segmentation." International Journal of Computer Vision **59**(2): 167-181.

Fleet, D. J., M. J. Black, et al. (2000). "Design and Use of Linear Models for Image Motion Analysis." International Journal of Computer Vision **36**(3): 171-193.

Flitton, G., T. Breckon, et al. (2010). Object Recognition using 3D SIFT in Complex CT Volumes. British Machine Vision Conference 2010. F. Labrosse, R. Zwiggelaar, Y. Liu and B. Tiddeman. Aberystwyth, BMVA Press**:** 11.1-11.12.

Foix, S., G. Alenya, et al. (2011). "Lock-in Time-of-Flight (ToF) Cameras: A Survey." Sensors Journal, IEEE **11**(9): 1917-1926.

Forsyth, D. A. and J. Ponce (2003). Chapter 4: Colour. Computer Vision: A Moden Approach. Indianapolis, Prentice Hall**:** 80-121.

Forsyth, D. A. and J. Ponce (2003). Chapter 9: Edge Detection. Computer Vision: A Moden Approach. Indianapolis, Prentice Hall**:** 238-260.

Forsyth, D. A. and J. Ponce (2003). Computer Vision: A Moden Approach. Indianapolis, Prentice Hall**:** 245-246.

Forsyth, D. A. and J. Ponce (2003). Section 5: Mid-level Vision. Computer Vision: A Moden Approach. Indianapolis, Prentice Hall**:** 433-468.

Fry, P. (2011, 16 Aug 2011). "How many cameras are there?", from https://www.cctvusergroup.com/art.php?art=94.

Fukelsheim, F. (1994). "The Three Sigma Rule." The American Statistician **48**(2): 88-91.

Gavrila, D. M. (1999). "The Visual Analysis of Human Movement: A Survey." Computer Vision and Image Understanding **73**(1): 82-98.

Giorgi, D., M. Mortara, et al. (2010). 3D Shape Retrieval Based on Best View Selection. Proceedings of the ACM workshop on 3D object retrieval. Firenze, Italy, ACM.

Gorelick, L., M. Blank, et al. (2007). "Actions as Space-Time Shapes." Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(12): 2247-2253.

Gorelick, L., M. Galun, et al. (2006). "Shape Representation and Classification Using the Poisson Equation." IEEE Trans. Pattern Anal. Mach. Intell. **28**(12): 1991-2005.

Grundmann, M., V. Kwatra, et al. (2010). Efficient Hierarchical Graph-Based Video Segmentation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010. San Francisco, USA.

Guler, S., J. A. Silverstein, et al. (2007). Stationary Objects in Multiple Object Tracking. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS).

Hopfield, J. J. (1988). "Artificial neural networks." Circuits and Devices Magazine, IEEE **4**(5): 3-10.

Horn, B. K. P. (1987). "Motion Fields are Hardly Ever Ambiguous." International Journal of Computer Vision **1**(3): 259-274.

Horn, B. K. P. and B. G. Schunck (1981). "Determining Optical Flow." Artificial Intelligence **17**(1-3): 185-203.

Hothorn, T., P. Buhlmann, et al. (2010). "Model-based Boosting 2.0." The Journal of Machine Learning Research **99**: 2109-2113.

Hu, M.-K. (1962). "Visual Pattern Recognition by Moment Invariants." IRE Transactions on Information Theory **8**(2): 179-187.

Inoue, H., T. Tachikawa, et al. (1992). Robot Vision System with a Correlation Chip for Real-time Tracking, Optical Flow and Depth Map Generation. 1992 IEEE International Conference on Robotics and Automation.

Jain, A. K. and F. Farrokhnia (1991). "Unsupervised Texture Segmentation using Gabor Filters." Pattern Recognition **24**(12): 1167-1186.

Jarrett, K. (2010). "YouTube: Online Video and Participatory Culture." Continuum **24**(2): 327-330.

Jianbo, S. and J. Malik (2000). "Normalized Cuts and Image Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8): 888-905.

Jiang, H., M. S. Drew, et al. (2006). Successive Convex Matching for Action Detection. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, . **2:** 1646-1653.

Jong-Sen, L. (1981). "Refined Filtering of Image Noise Using Local Statistics." Computer Graphics and Image Processing **15**(4): 380-389.

Kass, M., A. Witkin, et al. (1988). "Snakes: Active Contour Models." International Journal of Computer Vision **1**(4): 321-331.

Katsuragawa, S., K. Doi, et al. (1988). "Image Feature Analysis and Computer‐Aided Diagnosis in Digital Radiography: Detection and Characterization of Interstitial Lung Disease in Digital Chest Radiographs." Medical Physics **15**(3): 311-319.

Ke, Y., R. Sukthankar, et al. (2010). "Volumetric Features for Video Event Detection " International Journal of Computer Vision **88**(3): 339-362.

Kilambi, P., O. Masoud, et al. (2006). Crowd Analysis at Mass Transit Sites. Intelligent Transportation Systems Conference. Toronto, Ont.**:** 753-758.

Koller, D., J. Weber, et al. (1994). Towards robust Automatic Traffic Scene Analysis in Real-time. Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference onPattern Recognition.

Kuhne, G., S. Richter, et al. (2001). Motion-based segmentation and contour-based classification of video objects. Proceedings of the ninth ACM international conference on Multimedia. Ottawa, Canada, ACM.

Laptev, I. and T. Lindeberg (2003). Space-time Interest Points. 9th IEEE International Conference onComputer Vision, 2003. Proceedings.**:** 432-439 vol.1.

Lewis, J. P. (1995). "Fast Normalized Cross-correlation." Vision Interface. Canadian Image Processing and Pattern Recognition Society: 120-123.

Li, F. F. and P. Perona (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Li, Y., C. K. Tang, et al. (2001). Efficient Dense Depth Estimation from Dense Multi-Perspective Panoramas. Eighth IEEE International Conference on Computer Vision. Vancouver BC. **1:** 119-126.

Lindeberg, T. (1994). "Scale-space Theory: A Basic Tool for Analyzing Structures at Different Scales." Journal of Applied Statistics **21**(1-2): 225-270.

Liu, C. B. and N. Ahuja (2004). Vision Based Fire Dectection. 17th International Conference on Pattern Recognition. Cambridge, UK. **4:** 134-137.

Lopes, A. P. B., R. S. Sloveira, et al. (2009). Spatio-Temporal Frames in a Bag-of-visual-features Approach for Human Actions Recognition. 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing. Rio de Janiero**:** 315-321.

Lowe, D. G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints." International Journal of Computer Vision **60**(2): 91-110.

MacEvoy, B. (2010, 2010). "CIELUV uniform color space. 2010, from http://www.handprint.com/HP/WCL/color7.html#CIELUV.

Marin, V. H. (1987). "Angular Reconstitution: A Posteriori Assignment of Projection Directions for 3D Reconstruction." Ultramicroscopy **21**(2): 111-123.

Matikainen, P., M. Hebert, et al. (2009). Trajectons: Action recognition through the motion analysis of tracked features. 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops).

Mittal, A. and N. Paragios (2004). Motion-based Background Subtraction using Adaptive Kernel Density Estimation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Moeslund, T. B. and E. Granum (2001). "A Survey of Computer Vision-based Human Motion Capture." Computer Vision and Image Understanding **81**(3): 231-268.

Moeslund, T. B., A. Hilton, et al. (2006). "A Survey of Advances in Vision-based Human Motion Capture and Analysis." Computer Vision and Image Understanding **104**(2): 90-126.

Moghaddam, B. and A. Pentland (1997). "Probabilistic Visual learning for Object Representation." Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(7): 696-710.

Mori, S., H. Nishida, et al. (1999). Optical Character Recognition. New York, NY, USA, John Wiley & Sons, Inc.

Ngo, C. W., T. C. Pong, et al. (2003). "Motion Analysis and Segmentation through Spatio-Temporal Slice Processing." IEEE Transactions on Image Pricessing **12**(3): 341-355.

Niebles, J. C., H. Wang, et al. (2008). "Unsupervised Learning of Human Action Categories Using Spatial-TemporalWords." International Journal of Computer Vision **79**(3): 299-318.

Niyogi, S. A. and E. H. Adelson (1994). Analyzing and recognizing walking figures in XYT. IEEE Computer Society Conference on Computer Vision and Pattern Recognition**:** 469-474.

nVIDIA. "What is CUDA." 2011, from http://developer.nvidia.com/what-cuda.

Oikonomopoulos, A., I. Patras, et al. (2005). "Spatiotemporal Salient Points for Visual Recognition of Human Actions." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **36**(3): 710-719.

Papageorgiou, C. P., M. Oren, et al. (1998). A General Framework for Object Detection. Computer Vision, 1998. Sixth International Conference on.

Peng, H., Z. Ruan, et al. (2010). "V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets." Nature Biotechnology **28**(4): 348-353.

"People Falling Down Funny? The Ugly Face of Public Health Care." from http://www.youtube.com/watch?v=NKbtXtJLviU.

Popper, R. (2010). "A Survey on Vision-based Human Action Recognition." Image and Vision Cimputing **28**(6): 976-990.

Porikli, F. (2004). Learning Object Trajectory Patterns by Spectral Clustering. IEEE International Conference on Multimedia and Expo. Taipei. **2:** 1171-1174.

Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning **1**(1): 81-106.

Rav-Acha, A. and S. Peleg (2004). A Unified Approach for Motion Analysis and View Synthesis. 2nd International Symposium on 3D Data Processing, Visualization and transmission. **1:** 717-724.

Reng, L., T. B. Moeslund, et al. (2005). Finding Motion Primitives in Human Body Gestures. 6th International Gesture Workshop. S. Gibet, N. Courty and J. F. Kamp. Berder Island, France, Springer**:** 133-144.

Rui, Y., T. S. Huang, et al. (1999). "Image Retrieval: Current Techniques, Promising Directions, and Open Issues." Journal of Visual Communication and Image Representation **10**(1): 39-62.

Sayood, K. (1996). Data Compression. USA, Acadeic Press.

Schanda, J. (2007). 3. CIE Colorimetry. Colorimetry: Understanding the CIE System, Wiley**:** 43-44.

Schuldt, C., I. Laptev, et al. (2004). Recognising Human Actions: A Local SVM Approach. International Conference on Pattern Recognition. Cambridge, UK**:** 32-36.

Scott, D. W. (1992). Kernel Density Estimatiors. Multivariate Density Estimation: Theory, Practice and Visualization. New York, Wiley-Interscience Publication**:** 149-155.

Shi, J. and C. Tomasi (1994). Good Features to Track. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA**:** 593-600.

Siva, P. and T. Xiang (2010). Action Detection in Crowd. British Machine Vision Conference 2010. F. Labrosse, R. Zwiggelaar, Y. Liu and B. Tiddeman. Aberystwyth, UK, BMVA Press**:** 9.1-9.11.

Stone, Z., T. Zickler, et al. (2008). Autotagging Facebook: Social Network Context Improves Photo Annotation. Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on.

Suma, E. A., B. Lange, et al. (2011). FAAST: The Flexible Action and Articulated Skeleton Toolkit. Virtual Reality Conference (VR), 2011 IEEE.

Szeliski, R. (2010). Computer Vision: Algorithms and Applications. D. Gries. Ithaca, NY USA, Springer**:** 122-131.

Szeliski, R. (2010). Computer Vision: Algorithms and Applications. D. Gries. Ithaca, NY USA, Springer**:** 115-117.

Tangelder, J. and R. Veltkamp (2008). "A survey of content based 3D shape retrieval methods." Multimedia Tools and Applications **39**(3): 441-471.

Teague, M. R. (1980). "Image Analysis via the General Theory of Moments*." Journal of the Optical Society of America A **70**(8): 920-930.

Terzopoulos, D., A. Witkin, et al. (1988). "Constraints on Deformable Models:Recovering 3D Shape and Nonrigid Motion." Artificial Intelligence **36**(1): 91-123.

Toyama, K., J. Krumm, et al. (1999). Wallflower: Principles and Practice of Background Maintenance. The 7th IEEE International Conference on Computer Vision.

Viola, P. and M. J. Jones (2004). "Robust Real-Time Face Detection." International Journal of Computer Vision **57**(2): 137-154.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, ACM.

Wang, J. and M. F. Cohen (2007). "Image and Video Matting: A Survey." Found. Trends. Comput. Graph. Vis. **3**(2): 97-175.

Wang, J., Z. Xu, et al. (2010). Head Curve Matching and Graffiti Detection. British Machine Vision Conference 2010. Aberystwyth, UK.

Wang, L. and D. Suter (2007). "Learning and Matching of Dynamic Shape Manifolds for Human Action Recognition." IEEE Tansactions on Image Processing **16**(6): 1646-1661.

Wang, Y., K. F. Loe, et al. (2005). "Spatiotemporal video segmentation based on graphical models." Image Processing, IEEE Transactions on **14**(7): 937-947.

Wang, Y. and G. Mori (2009). "Human Action Recognition by Semilatent Topic Models." IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(10): 1762-1774.

Weinland, D., R. Ronfard, et al. (2006). "Free viewpoint action recognition using motion history volumes." Computer Vision and Image Understanding **104**(2-3): 249-257.

Weiss, Y. (1999). Segmentation using Eigenvectors: A Unifying View. The Proceedings of the 7th IEEE International Conference on Computer Vision.

Wieser, D. and T. Brupbacher (2001). Smoke Detection in Tunnels using Video Images. 12th International Conference on Automatic Detection. Gaithersburg, MD, USA**:** 79-90.

Willems, G., T. Tuytelaars, et al. (2008). "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector." Lecture Notes In Computer Science; Proceedings of the 10th European Conference on Computer Vision: Part II **5303**: 650-663.

Williams, L. (1983). "Pyramidal parametrics." SIGGRAPH Comput. Graph. **17**(3): 1-11.

Wright, J., A. Y. Yang, et al. (2009). "Robust Face Recognition via Sparse Representation." Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(2): 210-227.

Wu, Z. and R. Leahy (1993). "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(11): 1101-1113.

Yang, M. H. and N. Ahuja (1998). "Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases." Storage and Retrieval for Image and Video Databases VII **3656**: 458-466.

Yeung, M. M. and Y. Boon-Lock (1997). "Video visualization for compact presentation and fast browsing of pictorial content." Circuits and Systems for Video Technology, IEEE Transactions on **7**(5): 771-785.

Yuille, A. L. and N. M. Grzymacz (1988). "A Computational Theory for the Perception of Coherent Visual Motion." Nature **333**: 71-74.

Zhao, T. and R. Nevatia (2002). Stochastic Human segmentation from a Static Camera. Workshop on Motion and Video Computing 2002. Orlando, USA. **1:** 9-14.

Zhou, H. and H. Hu (2008). "Human Motion Tracking for Rehabilitation - A Survey." Biomedical Signal Processing and Control **3**(1): 1-18.

Zoran, A. and M. Coelho (2011). "Cornucopia: The Concept of Digital Gastronomy." Leonardo **44**(5): 425-431.